Statistics and Operations Research Transactions,

vol. 41, n. 1 (2017)

SORT 41 (1) January-June 2017, 3-54 DOI: 10.2436/20.8080.02.49

Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling

Anne Chao[1,*] and Robert K. Colwell[2,3,4]

Abstract

In the context of capture-recapture studies, Chao (1987) derived an inequality among capture frequency counts to obtain a lower bound for the size of a population based on individuals' capture/non-capture records for multiple capture occasions. The inequality has been applied to obtain a non-parametric lower bound of species richness of an assemblage based on species incidence (detection/non-detection) data in multiple sampling units. The inequality implies that the number of undetected species can be inferred from the species incidence frequency counts of the uniques (species detected in only one sampling unit) and duplicates (species detected in exactly two sampling units). In their pioneering paper, Colwell and Coddington (1994) gave the name "Chao2" to the estimator for the resulting species richness. (The "Chao1" estimator refers to a similar type of estimator based on species abundance data). Since then, the Chao2 estimator has been applied to many research fields and led to fruitful generalizations. Here, we first review Chao's inequality under various models and discuss some related statistical inference questions: (1) Under what conditions is the Chao2 estimator an unbiased point estimator? (2) How many additional sampling units are needed to detect any arbitrary proportion (including 100%) of the Chao2 estimate of asymptotic species richness? (3) Can other incidence frequency counts be used to obtain similar lower bounds? We then show how the Chao2 estimator can be also used to guide a non-asymptotic analysis in which species richness estimators can be compared for equally-large or equally-complete samples via sample-size-based and coverage-based rarefaction and extrapolation. We also review the generalization of Chao's inequality to estimate species richness under other sampling-without-replacement schemes (e.g. a set of quadrats, each surveyed only once), to obtain a lower bound of undetected species shared between two or multiple assemblages, and to allow inferences about undetected phylogenetic richness (the total length of undetected branches of a phylogenetic tree connecting all species), with associated rarefaction and extrapolation. A small empirical dataset for Australian birds is used for illustration, using online software SpadeR, iNEXT, and PhD.

MSC: 62D07, 62P07.

Keywords: Cauchy-Schwarz inequality, Chao2 estimator, extrapolation, Good-Turing frequency formula, incidence data, phylogenetic diversity, rarefaction, sampling effort, shared species richness, species richness.

[*] Corresponding author. E-mail: chao@stat.nthu.edu.tw

[1] Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan.

[2] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA.

[3] University of Colorado Museum of Natural History, Boulder, CO 80309, USA.

[4] Departmento de Ecologia, Universidade Federal de Goiás, CP 131, 74.001-970, Goiânia, GO, Brasil.

Received: December 2016.

1. Introduction

Thirty years ago, Chao (1987) developed an inequality among capture frequency counts to obtain a lower bound of population size based on individuals' capture/non-capture records in multiple-stage, closed capture-recapture studies. An earlier version of Chao's inequality and the corresponding lower bound (Chao, 1984) estimated the number of classes under a classic occupancy problem. Those inequalities and lower bounds were derived for their pure mathematical interest, as the models are simple and elegant, and also for their statistical interest, because these inequalities can be used to make inference about the richness of the undetected portion of a biological assemblage based on incomplete data.

In the first decade after their publication, these Chao-type lower bounds were rarely applied in other disciplines. In 1994, Colwell and Coddington published a seminal paper on estimating terrestrial biodiversity through extrapolation. They applied both of Chao's formulas (1984, 1987) to estimate species richness, because there is a simple analogy between the incidence data in species richness estimation for a multiple-species assemblage and the capture-recapture data in population size estimation for a single species. Chao (1984) had suggested that her occupancy-based estimator might be applied to estimating species richness, and offered examples of its application to capture-recapture data, the focus of Chao (1987). Colwell and Coddington distinguished two types of data: individual-based abundance data (counts of the number of individuals of each species within a single sampling unit) and multiple sampling-unit-based incidence data (counts of occurrences of each species among sampling units). They gave the name "Chao1" to the estimator of species richness specifically for abundance data, based on the Chao (1984) formula, and the name "Chao2" for incidence data based on the Chao (1987) formula. Colwell also featured these two estimators along with others in the widely used software EstimateS (Colwell, 2013; Colwell and Elsensohn, 2014). Since then, both the Chao1 and Chao2 estimators have been increasingly applied to many research fields, not only in ecology and conservation biology, but also in other disciplines; see Chazdon et al. (1998), Magurran (2004), Chao (2005), Gotelli and Colwell (2011), Magurran and McGill (2011), Gotelli and Chao (2013) and Chao and Chiu (2016) for various applications. Chao's inequalities also led to numerous generalizations under different models or frameworks; some closely related generalizations were accomplished by Mao (2006, 2008), Mao and Lindsay (2007), Rivest and Baillargeon (2007), Pan, Chao and Foissner (2009), Böhning and van der Heijden (2009), Lanumteangm and Böhning (2011), Böhning et al. (2013), Mao et al. (2013), Chiu et al. (2014), and Puig and Kokonendji (2017). In addition to EstimateS, these two estimators have now been included in other software and several R packages in CRAN (e.g. packages Species, Specpool, entropart, fossil, SpadeR, iNEXT, among others).

During the past 30 years, Chao and her students and collaborators have developed a number of population size and species richness estimators based on several other statistical models, including Chao and Lee's (1992) abundance- or incidence-based coverage estimators (ACE and ICE, two names bestowed by Chazdon et al., 1998), martingale estimators, estimating-function estimators, maximum quasi-likelihood estimators, and Horvitz-Thompson-type estimators; see Chao (2001) and Chao and Chiu (2016) for a review. These developments are more complicated and mathematically sophisticated than the estimators derived from Chao's inequalities. Surprisingly, it turns out that the earliest and simplest estimators are the most useful ones for biological applications.

In this paper, we mainly focus on Chao's (1987) inequality and its subsequent developments for multiple incidence data. For both practical and biological reasons, recording species detection/non-detection in multiple sampling units is often preferable to enumerating individuals in a single sampling unit (abundance data). For microbes, clonal plants, and sessile invertebrates, individuals are difficult or impossible to define. For mobile organisms, replicated incidence data are less likely to double-count individuals. For social animals, counting the individuals in a flock, herd, or school may be difficult or impractical. Also, replicated incidence data support statistical approaches to richness estimation that are just as powerful as corresponding abundance-based approaches (Chao et al., 2014b). Moreover, a further advantage is that replicated incidence records account for spatial (or temporal) heterogeneity in the data (Colwell et al., 2004, 2012).

In Sections 2.1 and 2.2, we first review the general model formulation for incidence data and the Chao (1987) inequality. Three related statistical inference problems are discussed:

1. In Section 2.3, we ask under what conditions the Chao2 estimator is an unbiased point estimator. Chao et al. (2017) recently provided an intuitive answer to this question for abundance data, from a Good-Turing perspective. Here we use a generalization of the Good-Turing frequency formula to answer the same question for incidence data.
2. In Section 2.4, we ask how many additional sampling units are needed to detect any arbitrary proportion (including 100%) of the Chao2 estimate. The Chao2 species richness estimator does not indicate how much sampling effort (additional sampling units) would be necessary to answer the question. Here we review the solution proposed by Chao et al. (2009).
3. In Section 2.5, we review approaches that use other incidence frequency counts to obtain similar-type lower bounds. In Chao's (1987) formula, the estimator for the number of undetected species is based only on the frequency counts of the uniques (species detected in only one sampling unit) and duplicates (species detected in exactly two sampling units). Lanumteangm and Böhning (2011), Chiu et al. (2014), Puig and Kokonendji (2017) made advances by extending Chao's inequality to use higher-order incidence frequency counts. Here we mainly review Puig and Kokonendji's (2017) extension, which leads to a series of lower bounds for species richness. Their framework was based mainly on abundance data, but it can be readily applied to multiple incidence data.

In Section 3, we show that, no matter whether the Chao2 formula is unbiased or biased low, it can always be used to guide a non-asymptotic analysis in which a species richness estimator can be compared for equally-large samples (based on a common number of sampling units) or equally-complete samples (based on a common value of sample completeness, as measured by coverage; see later text). Sample-size-based and coverage-based rarefaction and extrapolation provide a unified sampling approach to fairly comparing species richness across assemblages.

In the subsequent three sections we review three generalizations of Chao's inequality to estimate species richness under other sampling schemes (Section 4), to estimate shared species richness between two or multiple assemblages (Section 5), and also to make inferences about phylogenetic diversity, which incorporates species evolutionary history (Section 6). The next three paragraphs introduce these generalizations.

Chao's original inequality was developed under the assumption that sampling units are assessed with replacement. When sampling is done without replacement, e.g. quadrats or time periods are not repeatedly selected/surveyed, or mobile species are collected by lethal sampling methods, suitable modification is needed. In Section 4, we review the modifications developed by Chao and Lin (2012).

Compared with estimating species richness in a single assemblage, the estimation of shared species richness, taking undetected species into account, has received relatively little attention; see Chao and Chiu (2012) for a review. For two assemblages, shared species richness plays an important role in assessing assemblage overlap and forms a basis for constructing various types of beta diversity and (dis)similarity measures, such as the classic Sørensen and Jaccard indices (Colwell and Coddington, 1994; Magurran, 2004; Chao et al., 2005, 2006; Jost, Chao and Chazdon, 2011; Gotelli and Chao, 2013). In Section 5, we review the work by Pan et al. (2009), who extended Chao's inequality to the case of multiple assemblages to obtain a lower bound of undetected species shared between two or multiple assemblages.

A rapidly growing literature discusses phylogenetic diversity, which incorporates evolutionary histories among species into diversity analysis (see Faith, 1992; Warwick and Clarke, 1995; Crozier, 1997; Webb and Nonoghue, 2005; Petchey and Gaston, 2002; Cadotte et al., 2009; Cavender-Bares, Ackerly and Kozak, 2012). The most widely used phylogenetic metric is Faith's (1992) *PD* (phylogenetic diversity), which is defined as the sum of the branch lengths of a phylogenetic tree connecting all species in the target assemblage. As shown by Chao et al. (2010, 2015), *PD* can be regarded as a measure of phylogenetic richness, i.e. a phylogenetic generalization of species richness. Throughout this paper, *PD* refers to Faith's (1992) *PD*. When some species are present, but undetected by a sample, the lineages/branches associated with these undetected species are also missing from the phylogenetic tree spanned by the observed species. The undetected *PD* in an incomplete sample was not discussed until recent years (Cardoso et al., 2014; Chao et al., 2015). In Section 6, we review the phylogenetic version of Chao's inequality, developed recently by Chao et al. (2015), and the associated phylogenetic version of the rarefaction/extrapolation approach.

In Section 7, a small empirical dataset for Australian birds is used for illustration using online software, including Chao's SpadeR, iNEXT, and PhD. Section 8 provides discussion and conclusions. The diversity measures discussed in this review (species richness, shared species richness, and *PD*) do not take species abundances into account. We briefly discuss the extension of these measures to incorporate species abundances, and refer readers to relevant papers. Major notation used in each section is shown in Table 1.

2. Species richness estimation

2.1. A general framework: Sampling-unit-based incidence data and model

As indicated in the Introduction, Chao's (1987) original inequality was formulated based on a capture-recapture model to estimate the size of a population, but here we consider a framework based on species incidence (detection/non-detection) data to estimate species richness. These two statistical inference problems are equivalent. Assume that there are S species indexed $1,2,\dots,S$ in the focal assemblage, where S is the estimating target in species richness estimation. Here we mainly consider the model developed by Colwell et al. (2012) for multiple incidence data. Assume that there are Tsampling units, and that they are indexed $1,2,\dots,T$. The sampling unit is usually a trap, net, quadrat, plot, or timed survey, and it is these sampling units, not the individual organisms, that are sampled randomly and independently. The observed data consist of species detection/non-detection in each sampling unit. In a typical spatial study, these sampling units are deployed randomly in space within the area encompassing the assemblage. However, in a temporal study of diversity, the T sampling units would be deployed in one place at different independent points in time (such as an annual breeding bird census at a single site).

For any sampling unit, the model assumes that the ith species has its own unique incidence or detection probability π_i that is constant among all randomly selected sampling units. The incidence probability π_i is the probability that species i is detected in a sampling unit. Here $\sum_{i=1}^{S}\pi_i$ will generally not be equal to unity.

The incidence records consist of a species-by-sampling-unit incidence matrix $\{W_{ij};\ i=1,2,\dots,S,\ j=1,2,\dots,T\}$ with S rows and T columns; here $W_{ij}=1$ if species i is detected in sampling unit j, and $W_{ij}=0$ otherwise. Let Y_i be the number of sampling units in which species i is detected, $Y_i=\sum_{j=1}^{T}W_{ij}$; here Y_i is referred to as the sample *species incidence frequency*. Species present in the assemblage but not detected in any sampling unit yield $Y=0$. See Section 6.1 for a hypothetical example and Appendices A and B for real data. Details about these data are provided in subsequent sections.

Table 1: *Major notation used in each section.*

Common notation and/or one-assemblage species richness estimation (Section 2)	
S	Number of species in an assemblage.
π_i	Detection or incidence probability of species i, $i = 1, 2, \ldots, S$ in a sampling unit.
T	Number of sampling units taken from an assemblage.
U	Total number of incidences in T sampling units.
ϕ_r	Mean detection probability of species that appeared in r sampling units, $r = 0, 1, \ldots, T$.
W_{ij}	Species detection/non-detection: $W_{ij} = 1$ if species i is detected in sampling unit j, and $W_{ij} = 0$ otherwise, $i = 1, 2, \ldots, S$, $j = 1, 2, \ldots, T$.
S_{obs}	Number of observed species in T sampling units.
Y_i	Species incidence frequency (number of sampling units in which species i is detected).
Q_k	Number of species detected in exactly k sampling units in the data, $k = 0, 1, \ldots, T$.
ˆ	"Hat" above a parameter: an estimator of the parameter, e.g. $\hat{S}$, $\hat{\pi}_i$ and $\hat{\phi}_r$ denote, respectively, estimators of S, π_i and ϕ_r.
Rarefaction and extrapolation of one-assemblage species richness (Section 3)	
$C(T)$	Coverage for a reference sample of size T.
$C(t)$	Coverage in a hypothetical rarefied sample of t sampling units if $t < T$.
$C(T+t^*)$	Coverage in a hypothetical augmented sample of $T+t^*$ sampling units.
$S(t)$	Expected number of species in a hypothetical rarefied sample of t sampling units if $t < T$.
$S(T+t^*)$	Expected number of species in a hypothetical augmented sample of $T+t^*$ sampling units.
One-assemblage species richness under sampling without replacement (Section 4)	
T^*	Total number of sampling units in the entire assemblage (e.g. total number of disjoint, equal-area quadrats in a region).
U_i	Number of sampling units (or quadrats) that species i can be detected.
q	Known sampling fraction, $q = T/T^*$.
Two-assemblage shared species richness estimation (Section 5)	
S_{12}	Number of shared species between Assemblages I and II.
π_{i1}, π_{i2}	Detection or incidence probability of species i, $i = 1, 2, \ldots, S$, in any sampling unit taken, respectively, from Assemblages I and II.
T_1, T_2	Number of sampling units in Samples I and II taken, respectively, from Assemblages I and II.
Y_{i1}, Y_{i2}	Species incidence frequency (i.e. number of sampling units in which species i is detected), respectively, in Samples I and II.
Q_{rv}	Number of shared species that are detected in r sampling units in Sample I and v sampling units in Sample II, r, $v = 0, 1, 2, \ldots$.
Q_{r+}	Number of shared species that are detected in r sampling units in Sample I and that are detected in at least one sampling unit in Sample II, $r = 0, 1, 2, \ldots, T_1$.
Q_{+v}	Number of shared species that are detected in v sampling units in Sample II and that are detected in at least one sampling unit in Sample I, $v = 0, 1, 2, \ldots, T_2$.
Q_{++}	Total number of observed species shared between Samples I and II.

Table 1 (cont.): *Major notation used in each section.*

One-assemblage phylogenetic diversity (*PD*) Estimation (Section 6)	
B	Number of branches/nodes in the phylogenetic tree spanned by all species of an assemblage.
L_i	Length of the ith branch/node.
PD	Sum of branch lengths in a phylogenetic tree.
λ_i	Detection or incidence probability of branch/node i, i.e. the probability of detecting at least one species descended from branch/node i in a sampling unit.
W^*_{ij}	Node detection/non-detection: $W^*_{ij} = 1$ if at least one species descended from branch i is detected in jth sampling unit, and $W^*_{ij} = 0$ otherwise, $i = 1,2,\ldots,B$, $j = 1,2,\ldots,T$.
PD_{obs}	PD in the observed tree.
Y^*_i	branch/node incidence frequency for branch/node i, $i = 1,2,\ldots,B$.
R_k	Sum of branch lengths for the branches with node incidence frequency $= k$, $k = 0,1,\ldots,T$.
Q^*_1, Q^*_2	Number of nodes/branches with incidence frequency $= 1$ and $= 2$, respectively, in the observed tree.
Rarefaction and extrapolation of one-assemblage *PD* (Section 6 and Table 2)	
$PD(t)$	Expected PD in a hypothetical rarefied sample of t sampling units if $t < T$.
$PD(T+t^*)$	Expected PD in a hypothetical augmented sample of $T+t^*$ sampling units.

Following Colwell et al. (2012), we assume, given the set of detection probabilities $(\pi_1, \pi_2, \ldots, \pi_S)$, that each element W_{ij} in the incidence matrix is a Bernoulli random variable with probability π_i. The probability distribution for the incidence matrix can be expressed as

$$P(W_{ij} = w_{ij};\ i = 1,2,\ldots,S,\ j = 1,2,\ldots,T) = \prod_{j=1}^{T}\prod_{i=1}^{S} \pi_i^{w_{ij}}(1-\pi_i)^{1-w_{ij}}$$
$$= \prod_{i=1}^{S} \pi_i^{y_i}(1-\pi_i)^{T-y_i}. \quad (1a)$$

The marginal distribution for the incidence-based frequency Y_i for the i-th species follows a binomial distribution characterized by T and the detection probability π_i:

$$P(Y_i = y_i) = \binom{T}{y_i} \pi_i^{y_i}(1-\pi_i)^{T-y_i}, \quad i = 1,2,\ldots,S. \quad (1b)$$

Denote the *incidence frequency counts* by $(Q_1, Q_2, \ldots, Q_T)$, where Q_k is the number of species detected in exactly k sampling units in the data, $k = 0,1,\ldots,T$. Here, Q_1 represents the number of "unique" species (those that are detected in only one sampling unit), and Q_2 represents the number of "duplicate" species (those that are detected in exactly two sampling units). The unobservable zero frequency count Q_0 denotes the number of species among the S species present in the assemblage that are not detected in any of the T sampling units. Then the number of observed species in the sample is $S_{obs} = \sum_{i>0} Q_i$ and $S_{obs} + Q_0 = S$.

2.2. Chao's inequality

Treating the incidence probabilities $(\pi_1, \pi_2, \ldots, \pi_S)$ as fixed, unknown parameters, we first present Chao's (1987) inequality under the model (1a) or (1b). Note the following expected value for the incidence frequency count Q_k:

$$E(Q_k) = E\left[\sum_{i=1}^{S} I(Y_i = k)\right] = \sum_{i=1}^{S} \binom{T}{k} \pi_i^k (1-\pi_i)^{T-k}, \quad k = 0,1,2,\ldots,T, \tag{1c}$$

where $I(A)$ is the indicator function, i.e. $I(A) = 1$ if the event A occurs, and is 0 otherwise. In particular, the expected number of undetected species, uniques and duplicates are respectively:

$$E(Q_0) = \sum_{i=1}^{S} (1-\pi_i)^T,$$

$$E(Q_1) = \sum_{i=1}^{S} T\pi_i (1-\pi_i)^{T-1},$$

$$E(Q_2) = \sum_{i=1}^{S} \binom{T}{2} \pi_i^2 (1-\pi_i)^{T-2}.$$

Chao (1987) proposed a lower bound of $E(Q_0)$ based on the following Cauchy-Schwarz inequality:

$$\left[\sum_{i=1}^{S} (1-\pi_i)^T\right] \left[\sum_{i=1}^{S} \pi_i^2 (1-\pi_i)^{T-2}\right] \geq \left[\sum_{i=1}^{S} \pi_i (1-\pi_i)^{T-1}\right]^2, \tag{2a}$$

equivalently,

$$E(Q_0) \times \frac{E(Q_2)}{\binom{T}{2}} \geq \left(\frac{E(Q_1)}{T}\right)^2.$$

Thus, a theoretical lower bound for $E(Q_0)$ is derived as

$$E(Q_0) \geq \frac{(T-1)}{T} \frac{[E(Q_1)]^2}{2E(Q_2)},$$

implying a theoretical lower bound for species richness:

$$S = E(S_{obs}) + E(Q_0) \geq E(S_{obs}) + \frac{(T-1)}{T} \frac{[E(Q_1)]^2}{2E(Q_2)}.$$

Replacing the expected values in the above with the observed data, we then obtain an estimated lower bound of species richness, with a slight modification when $Q_2 = 0$ (Colwell and Coddington, 1994, gave the name *Chao*2 to this estimator):

$$\hat{S}_{Chao2} = \begin{cases} S_{obs} + \dfrac{(T-1)}{T}\dfrac{Q_1^2}{2Q_2}, & \text{if } Q_2 > 0, \\ S_{obs} + \dfrac{(T-1)}{T}\dfrac{Q_1(Q_1-1)}{2}, & \text{if } Q_2 = 0. \end{cases} \tag{2b}$$

The estimated number of undetected species is based exclusively on the information on the least frequent species (the number of uniques and duplicates). This is based on a basic concept that the frequent/abundant species (those that occur in many sampling units) carry negligible information about the undetected species; only rare/infrequent species carry such information.

When does the Chao2 formula provide a nearly unbiased estimator? The Cauchy-Schwarz inequality in Eq. (2a) becomes an equality if and only if the species detection probabilities are homogeneous, that is, $\pi_1 = \pi_2 = \cdots = \pi_S$. Homogeneity of detection probabilities would be a very restrictive condition, one that is almost never satisfied in most practical applications, such as species abundance or incidence distributions in nature. However, as we will show in Section 2.3, this condition can be considerably relaxed from a different derivation/perspective. Note that in Chao's inequality (2a), only three expected frequency counts are involved: $E(Q_0)$, $E(Q_1)$ and $E(Q_2)$. The frequent species (species with relatively large detection probabilities) would tend to occur in many sampling units and thus generally do not contribute to any of these three terms. On the other hand, only rare/infrequent species (species with relatively low detection probabilities) would either be undetected or detected in only one or two sampling units and thus are those species that contribute to the three terms. Therefore, a relaxed condition for an unbiased Chao2 estimator is that *very rare/infrequent* species have approximately the same detection probabilities, and frequent species are allowed to be highly heterogeneous without affecting the estimates. A more rigorous justification is given in Section 2.3.

Applying a standard asymptotic approach (Chao, 1987), the following estimated variance estimators can be obtained if Q_1, $Q_2 > 0$:

$$\widehat{\text{var}}(\hat{S}_{Chao2}) = Q_2\left[\frac{1}{4}\left(\frac{T-1}{T}\right)^2\left(\frac{Q_1}{Q_2}\right)^4 + \left(\frac{T-1}{T}\right)^2\left(\frac{Q_1}{Q_2}\right)^3 + \frac{1}{2}\left(\frac{T-1}{T}\right)\left(\frac{Q_1}{Q_2}\right)^2\right], \tag{3a}$$

If $Q_1 > 0$, $Q_2 = 0$, the variance becomes

$$\widehat{\text{var}}(\hat{S}_{Chao2}) = \frac{1}{4}\left(\frac{T-1}{T}\right)^2 Q_1(2Q_1-1)^2 + \frac{1}{2}\left(\frac{T-1}{T}\right) Q_1(Q_1-1) - \frac{1}{4}\left(\frac{T-1}{T}\right)^2 \frac{Q_1^4}{\hat{S}_{Chao2}}. \tag{3b}$$

In the special case that $Q_1 = 0$, we have $\hat{S}_{Chao2} = S_{obs}$, implying that sampling is complete and there are no undetected species in the data; an approximate variance of S_{obs} can be obtained using an analytic method (Colwell, 2013) or a bootstrap method (see Section 3.3). When $Q_1 > 0$ so that $\hat{S}_{Chao2} > S_{obs}$, the distribution of $\hat{S}_{Chao2} - S_{obs}$ is generally skewed to the right. Using a log-transformation by treating $\log(\hat{S}_{Chao2} - S_{obs})$ as an approximately normal random variable, we obtain a 95% confidence interval for S: (Chao, 1987)

$$[S_{obs} + (\hat{S}_{Chao2} - S_{obs})/R,\ \ S_{obs} + (\hat{S}_{Chao2} - S_{obs})R], \tag{3c}$$

where $R = \exp\{1.96[\log(1 + \widehat{\text{var}}(\hat{S}_{Chao2})/(\hat{S}_{Chao2} - S_{obs})^2)]^{1/2}\}$. In this case, the resulting lower confidence limit is always greater than or equal to the observed species richness, a sensible result.

The Chao2 estimator is also valid in a binomial-mixture model in which incidence probabilities $(\pi_1, \pi_2, \ldots, \pi_S)$ are assumed to be a random sample from an unknown distribution with density $h(\pi)$. Under this model, we have

$$E(Q_k) = S \int_0^1 \binom{T}{k} \pi^k (1-\pi)^{T-k} h(\pi)\, d\pi, \quad k = 0, 1, 2, \ldots T. \tag{4a}$$

The summation terms in the Cauchy-Schwarz inequality (2a) are replaced by integral terms:

$$\left[\int_0^1 (1-\pi)^T h(\pi) d\pi\right] \left[\int_0^1 \pi^2 (1-\pi)^{T-2} h(\pi) d\pi\right] \geq \left[\int_0^1 \pi(1-\pi)^{T-1} h(\pi) d\pi\right]^2. \tag{4b}$$

The above two formulas also lead to the same Chao2 formula given in Eq. (2b). In the special case that $h(\pi)$ is a beta distribution with parameters α and β, the resulting expected incidence-frequency count $\{E(Q_k), k = 0, 1, 2, \ldots, n\}$ correspond to the probabilities of a beta-binomial distribution. Under the two conditions (i) T is large and π is small, such that $T\pi$ tends to a positive constant, and (ii) β/T tends to a positive constant c, Skellam (1948) proved that $E(Q_k)$ tends to $(\alpha + k - 1)![(\alpha - 1)!k!]^{-1}[1/(1 + c)]^k[c/(1+c)]^\alpha$, which is the probability of a negative binomial variable taking the value k. This result theoretically justifies the inference that Chao's inequality is also valid for beta-binomial and negative binomial distributions. It is well known that beta-binomial and negative binomial can be used to describe spatially clustered (if sampling units are quadrats in an area) or temporally aggregated (if sampling units are different times) pattern of species; see Hughes and Madden (1993) and Shiyomi, Takahashi and Yoshimura (2000). Therefore, even though there is spatial/temporal heterogeneity pattern for species incidences, the lower bound and the associated estimation are still valid.

2.3. When is the Chao2 estimator nearly unbiased?

Alan Turing and I. J. Good, in their famous cryptanalysis to crack German ciphers during World War II, developed novel statistical methods to estimate the true frequencies of rare code elements (including still-undetected code elements), based on the observed frequencies in "samples" of intercepted Nazi code. After the War, Turing gave permission to Good to publish their statistical work. An influential paper by Good (1953) and one by Good and Toulmin (1956) presented Turing's wartime statistical work on the frequency formula and related topics; see Good (1983, 2000) for more details. The frequency formula is now referred to as the Good-Turing frequency formula, which has a wide range of applications in biological sciences, statistics, computer sciences, information sciences, and linguistics, among others (McGrayne, 2011, p. 100).

In an ecological context, Turing's statistical problem can be formulated as an estimation of the true frequencies of rare species when a random sample of individuals is drawn from an assemblage. In Turing's case, there were almost infinitely many rare species so that all samples have undetected species. The Good-Turing formula answers the following question: given a species that appears r times ($r = 0, 1, 2, \dots$) in a sample of n individuals that fails to detect all species present, what is its true relative frequency in the entire assemblage? Turing and Good focussed on the case of small r, i.e. rare species. Turing gave a surprisingly simple and remarkably effective answer that is contrary to most people's intuition; see Chao et al. (2017) for a review.

The Good-Turing original frequency formula was based on abundance data. We here extend their formula to incidence data to answer the following question: Given species incidence data of T sampling units, for those species that appeared in r ($r = 0, 1, 2, \dots$) out of T sampling units, what is the mean detection probability of species that appeared in r sampling units, ϕ_r? Such a mean detection probability can be mathematically expressed as

$$\phi_r = \sum_{i=1}^{S} \pi_i I(Y_i = r)/Q_r, \quad r = 0, 1, 2, \dots \tag{5a}$$

The numerator in Eq. (5a) represents the total incidence probabilities of those species that appeared in r sampling units. Dividing the total by Q_r, we obtain the mean detection probability per species, among those that each appeared in r sampling units. Note that, for the special case of $r = 0$, Eq. (5a) implies

$$\phi_0 Q_0 = \sum_{i=1}^{S} \pi_i I(Y_i = 0), \tag{5b}$$

which is the total detection probabilities of the undetected species. If one additional sampling unit can be added, then we can interpret it as the expected number of species in the additional sampling unit that are undetected in the original sample.

Here we derive the corresponding Good-Turing incidence frequency formula for multiple incidence data by treating $(\pi_1, \pi_2, \dots, \pi_S)$ as fixed, unknown parameters, although a similar derivation is also valid for binomial-mixture models. Under the model (Eq. 1b), in which the incidence frequencies Y_i, $i = 1, 2, \ldots, S$, follow a binomial distribution characterized by T and detection probability π_i, we can express the sum of the odds of π_i for those species that each appeared in r sampling units as follows:

$$\begin{aligned}
E\left[\sum_{i=1}^{S} \frac{\pi_i}{1-\pi_i} I(Y_i = r)\right] &= \sum_{i=1}^{S} \frac{\pi_i}{1-\pi_i} \binom{T}{r} \pi_i^r (1-\pi_i)^{T-r} \\
&= \sum_{i=1}^{S} \binom{T}{r} \pi_i^{r+1} (1-\pi_i)^{T-(r+1)} \\
&= \frac{\binom{T}{r}}{\binom{T}{r+1}} \left[\sum_{i=1}^{S} \binom{T}{r+1} \pi_i^{r+1} (1-\pi_i)^{T-(r+1)}\right] \\
&= \frac{(r+1)}{(T-r)} E(Q_{r+1}). \qquad (5c)
\end{aligned}$$

Assume that all species that appeared in r sampling units have approximately the same incidence probabilities. Then we have the following approximation formula:

$$E\left[\sum_{i=1}^{S} \frac{\pi_i}{1-\pi_i} I(Y_i = r)\right] \approx Q_r \frac{\phi_r}{1-\phi_r}.$$

Thus, ϕ_r can be obtained by solving the equation: $Q_r\phi_r/(1-\phi_r) \approx (r+1)Q_{r+1}/(T-r)$, based on Eq. (5c). We then obtain the corresponding Good-Turing formula for incidence data:

$$\hat{\phi}_r = \frac{(r+1)Q_{r+1}}{(T-r)Q_r + (r+1)Q_{r+1}} \approx \frac{(r+1)Q_{r+1}}{(T-r)Q_r}. \qquad (5d)$$

The original Good-Turing frequency formula for abundance data has a similar form as the above approximation, but with incidence frequency counts being replaced by abundance frequency counts.

Good (1983, p. 28) provided an intuitive justification for the abundance-based Good-Turing frequency formula. Here we follow Good's approach to give a similar justification for incidence data. Given an original sample, consisting of T sampling units, suppose one additional sampling unit can be added. We ask how many species that had appeared r times in the original sample would occur in the additional sampling unit. Based on Eq. (5a), the answer is simply $\sum_{i=1}^{S} \pi_i I(Y_i = r) = \phi_r Q_r$, which can be estimated by $(r+1)Q_{r+1}/(T-r)$ using the following simple reasoning. Notice that any species that appeared r times in the original sample and also occurs in the additional sampling unit must occur in $r+1$ sampling units in the enlarged sample consisting of $T+1$ sampling units. Then the total number of incidences of such species is $(r+1)Q_{r+1}$. Because the order in which sampling units were taken is assumed to be irrelevant, the average number of such species occurring in a single sampling unit is thus $(r+1)Q_{r+1}/(T+1)$, which is approximately equal to $(r+1)Q_{r+1}/(T-r)$ if r is small. Dividing this ratio by the number of such species, Q_r, we obtain the incidence-data-based Good-Turing frequency formula for ϕ_r as given in Eq. (5d).

For the special cases of $r = 0$ and $r = 1$, Eqs. (5b) and (5d) lead to

$$\widehat{\phi_0 Q_0} = \frac{Q_1}{T}, \hat{\phi}_1 = \frac{2Q_2}{(T-1)Q_1},$$

where $\widehat{\phi_0 Q_0}$ denotes the estimate of the product of ϕ_0 and Q_0. Intuitively, we expect that the mean incidence probability of all undetected species should not be more than that of all uniques in the sample, i.e. $\phi_0 \le \phi_1$, and this ordering is preserved by the corresponding estimates. Then we obtain the Chao2 lower bound for the number of undetected species by the following inequality:

$$\hat{Q}_0 = \frac{\widehat{\phi_0 Q_0}}{\hat{\phi}_0} \ge \frac{\widehat{\phi_0 Q_0}}{\hat{\phi}_1} = \frac{\frac{Q_1}{T}}{\frac{2Q_2}{(T-1)Q_1}} = \frac{(T-1)}{T}\frac{Q_1^2}{2Q_2}. \tag{5e}$$

Notice that, in the above derivation, if $\hat{\phi}_0 \approx \hat{\phi}_1$, then the inequality sign in Eq. (5e) becomes an equality sign. Therefore, from the Good-Turing perspective, the Chao2 lower bound is a nearly unbiased point estimator if all undetected and unique species in samples have the same mean detection probabilities. Such a conclusion is valid if very rare/infrequent species have approximately homogenous detection probabilities in any sampling unit (because this implies $\hat{\phi}_0 \approx \hat{\phi}_1$); in this case, frequent species could be highly heterogeneous without affecting the estimator.

2.4. How many sampling units are needed to reach the Chao2 estimate?

As discussed earlier, the Chao2 formula (in Eq. 2b) implies that sampling is complete when all species have been found in at least two sampling units, i.e. $Q_1 = 0$; in such a case, the estimated undetected species richness is 0 and the estimated species richness reduces simply to the observed number of species. This result also reveals that, whenever at least one species is found in only one sample $(Q_1 > 0)$, sampling is not complete and some species remain undetected. However, the Chao2 species richness estimator does not indicate how much sampling effort (how many additional sampling units) would be necessary to reach the Chao2 estimate (i.e. the first point at which there are no longer any singletons).

For incidence data, “sample size” means the number of sampling units. Chao et al. (2009) developed a non-parametric method for estimating the minimum sample size required to detect any arbitrary proportion (including 100%) of the estimated Chao2 species richness based on the Good-Turing formula discussed in Section 2.3. When the target is the Chao2 estimate, Chao et al. (2009) approach is to predict the minimum sample size t to achieve the following stopping rule: there are no uniques in the enlarged sample of size $T+t$, or equivalently, the expected number of uniques in the enlarged sample of size $T+t$ is less than 0.5, because the theoretical expected value may not be an integer.

Note that the number of uniques in the enlarged sample of size $T+t$ includes two groups of species: (1) any species observed in only one sampling unit in the original sample (i.e. those species with $Y_i = 1$) for which no additional incidences are detected in the additional t samples with probability $(1-\pi_i)^t$, and (2) any species not detected in the original sample (i.e. those species with $Y_i = 0$) for which detection in exactly one sampling unit is observed in the additional t sampling units with probability $t\pi_i(1-\pi_i)^{t-1}$. That is, the expected number of uniques in the enlarged $T+t$ sampling units is:

$$\sum\nolimits_{i=1}^{S}(1-\pi_i)^t I(Y_i=1)+\sum\nolimits_{i=1}^{S} t\pi_i(1-\pi_i)^{t-1} I(Y_i=0).$$

As discussed in Section 2.3, we assume that all uniques in the original sample have mean detection probability ϕ_1, and all previously undetected species have mean detection probability ϕ_0. Then the number of uniques in the enlarged $T+t$ sampling units will decline to < 0.5 when t satisfies

$$Q_1(1-\phi_1)^t + Q_0\, t\phi_0(1-\phi_0)^{t-1} < 0.5.$$

When we apply the Good-Turing incidence frequency formula to this equation, and substitute ϕ_1, ϕ_0 and Q_0 by $\hat{\phi}_1 = 2Q_2/[2Q_2+(T-1)Q_1]$, $\hat{\phi}_0 = Q_1/[Q_1+T\hat{Q}_0]$ and $\hat{Q}_0 = (1-1/T)Q_1^2/(2Q_2)$, then the required t must satisfy the following equation:

$$Q_1\left(1+\frac{t}{T}\right)\left[1-\frac{2Q_2}{(T-1)Q_1+2Q_2}\right]^t < 0.5.$$

The additional number of sampling units needed to reach the Chao2 estimate is approximately equal to $t = Tx^*$, where x^* is the solution of the following equation:

$$2Q_1(1+x) = \exp\left[x\frac{2Q_2}{(1-1/T)Q_1+2Q_2/T}\right]. \tag{6a}$$

If g is the fraction of $\hat{S}_{Chao2}$ that is desired ($0 < g < 1$), then the objective is to find the number of additional m_g sampling units such that the number of species reaches the target value $g\hat{S}_{Chao2}$, i.e. the expected number of previously undetected species that will be discovered in the additional m_g sampling units is $g\hat{S}_{Chao2} - S_{obs}$. This expected number, given the observed data, is

$$\sum_{i=1}^{S}[1-(1-\pi_i)^{m_g}]I(Y_i=0)\approx Q_0[1-(1-\phi_0)^{m_g}]. \tag{6b}$$

Applying the Good-Turing incidence frequency formula and substituting ϕ_0 and Q_0, we obtain that the required number of additional sampling units to reach a fraction g of $\hat{S}_{Chao2}$ (if $g\hat{S}_{Chao2}>S_{obs}$) is the number m_g such that $\hat{Q}_0[1-(1-\hat{\phi}_0)^{m_g}]=g\hat{S}_{Chao2}-S_{obs}$, i.e.

$$m_g\approx\frac{\log\left[1-\dfrac{T}{(T-1)}\dfrac{2Q_2}{Q_1^2}(g\hat{S}_{Chao2}-S_{obs})\right]}{\log\left[1-\dfrac{2Q_2}{(T-1)Q_1+2Q_2}\right]}. \tag{6c}$$

Chao et al. (2009) also provided an Excel spreadsheet for calculating necessary sampling effort for either abundance data or replicated incidence data.

2.5. A class of lower bounds

In the Chao2 approach (Eq. 2b), the estimator for undetected species richness is only in terms of the species incidence frequency counts of the uniques and duplicates in data. Several authors extended this approach to higher-order incidence frequency counts. Lanumteang and Böhning (2011) proposed using an additional incidence frequency count, i.e. the number of species that are detected in exactly three sampling units. They applied the above estimator to a variety of real data sets and concluded that the new estimator is especially useful for large populations and heterogeneous detection probabilities.

When the Chao2 estimator only provides a lower bound, its bias can be evaluated and assessed by using the Good-Turing frequency formula. In this case, an improved reduced-bias lower bound, which makes use of the additional information of Q_3 and Q_4, was derived by Chiu et al. (2014). The corresponding lower bound of species richness is referred to as *iChao2 estimator* (here the sub-index*i* stands for "improved"):

$$\hat{S}_{iChao2}=\hat{S}_{Chao2}+\frac{(T-3)}{4T}\frac{Q_3}{Q_4}\times\max\left(Q_1-\frac{(T-3)}{2(T-1)}\frac{Q_2Q_3}{Q_4},0\right). \tag{6d}$$

They also provided an analytic variance estimator to construct the associated confidence intervals.

Puig and Kokonendji (2017) extended Chao's inequality to a broader class of distributions that have log-convex probability generating functions. They obtained a series of lower bounds for the undetected species richness. This class of distribution includes compound Poisson distribution and Poisson-mixture distributions. Their framework is mainly based on abundance data, but it can be readily applied to multiple incidence data, as shown below.

Following the proof of Puig and Kokonendji (2017), we assume that the incidence probabilities $(\pi_1, \pi_2, \ldots, \pi_S)$ are a random sample from an unknown distribution with density $h(\pi)$, and we have $E(Q_k)$ given in Eq. (4a). Consider a probability density function:

$$H(\pi) = \frac{(1-\pi)^T h(\pi) d\pi}{\int_0^1 (1-u)^T h(u) du}, \quad 0 < \pi < 1.$$

Puig and Kokonendji (2017) showed the following moment inequality for $r, v = 0, 1, 2, \ldots$

$$\int_0^1 \left(\frac{\pi}{1-\pi}\right)^{r+v} H(\pi) d\pi \geq \int_0^1 \left(\frac{\pi}{1-\pi}\right)^{r} H(\pi) d\pi \times \int_0^1 \left(\frac{\pi}{1-\pi}\right)^{v} H(\pi) d\pi,$$

equivalently,

$$\left[\int_0^1 (1-\pi)^T h(\pi) d\pi\right] \left[\int_0^1 \pi^{r+v} (1-\pi)^{T-(r+v)} h(\pi) d\pi\right]$$
$$\geq \left[\int_0^1 \pi^{r} (1-\pi)^{T-r} h(\pi) d\pi\right] \left[\int_0^1 \pi^{v} (1-\pi)^{T-v} h(\pi) d\pi\right].$$

Then we have

$$E(Q_0) \geq \frac{\binom{T}{r+v} E(Q_r) \times E(Q_v)}{\binom{T}{r}\binom{T}{v} E(Q_{r+v})}, \quad r, v = 0, 1, 2, \ldots \tag{6e}$$

A series of lower bounds of S can then be obtained if $Q_{r+v} > 0$:

$$S_{obs} + \frac{\binom{T}{r+v} Q_r \times Q_v}{\binom{T}{r}\binom{T}{v} Q_{r+v}}, \quad r, v = 1, 2, \ldots$$

In the special case of $r = v = 1$, the above lower bound reduces to the Chao2 estimator. Puig and Kokonendji (2017) proved that, under a Poisson-mixture model, the greatest lower bound attains at the special case $r = v = 1$. This also provides a justification for the use of the Chao2 lower bound.

3. Species richness estimation for standardized samples: non-asymptotic analysis

Species richness estimation represents an “asymptotic” analysis; here “asymptotic” means that, as sample size tends to infinity, sample completeness approaches unity. When the Chao2 estimates are nearly unbiased under the conditions given in Section 2.3, they can be compared across multiple assemblages. However, when rare/infrequent species are highly heterogeneous and sample size is not sufficiently large, the Chao2 formula can provide only a lower bound, which cannot be compared accurately across assemblages, because the data provide insufficient information to accurately estimate species richness due to high heterogeneity of infrequent species. No matter whether or not Chao2 is unbiased, in any particular case, we can always use it to perform “non-asymptotic” analysis, in which samples are standardized based on a common finite sample size or on sample completeness via rarefaction and extrapolation. Again for incidence data, sample size refers to the number of sampling units.

The objective of a non-asymptotic approach is to control the dependence of the empirical species counts on sampling effort and sample completeness. The earliest development of standardization of sample size for abundance data by rarefaction was proposed by Sanders (1968), but see Chiarucci et al. (2008) for a historical review. Subsequent developments include studies by Hurlbert (1971), Simberloff (1972), Heck, van Belle and Simberloff (1975) and Coleman et al. (1982); see Gotelli and Colwell (2001, 2011) for details. Ecologists typically use rarefaction to down-sample the larger samples until they are the same size as the smallest sample. Ecologists then compare richness of these equally-large samples, but this approach implies that some data in larger samples are thrown away. To avoid discarding data, Colwell et al. (2012) proposed using a unified sample-size-based rarefaction (interpolation) and extrapolation (prediction) sampling curve for species richness, that can be rarefied to smaller sample sizes or extrapolated to larger sample sizes.

Chao and Jost (2012) indicated that a sample of a given size may be sufficient to fully characterize a low-diversity assemblage, but insufficient to characterize a rich-assemblage. Thus, when the species counts of two equally-large samples are compared, one might be comparing a nearly complete sample to a very incomplete one. In this case, any difference in diversity between the sites will generally be underestimated. They proposed rarefaction and extrapolation to a comparable degree of sample completeness (as measured by sample coverage; see below) and developed a coverage-based rarefaction and extrapolation methodology. The sample-size-based and coverage-based integration of rarefaction and extrapolation of species richness represent a unified sampling framework for quantifying and comparing species richness across multiple assemblages.

Here we review the sample-size-based and coverage-based rarefaction and extrapolation of species richness; all formulas are tabulated in the first and the third columns of Table 2.

Table 2: *The theoretical formulas and analytic estimators for rarefaction and extrapolation of species richness (left column), Faith's PD (middle column), and sample coverage (right column) based on incidence data, given a reference sample with observed species richness = S_{obs}, observed PD = PD_{obs}, and estimated coverage $\hat{C}(T)$ for incidence data. Here the sample size means the number of sampling units. See Colwell et al. (2012) and Chao and Jost (2012) for derivation details.*

Species richness	Faith's PD	Coverage
(a) Theoretical formula for any hypothetical sample size of t		
$S(t)=\sum_{i=1}^{S}[1-(1-\pi_i)^t]$	$PD(t)=\sum_{i=1}^{B}L_i[1-(1-\lambda_i)^t]$	$C(t)=1-\dfrac{\sum_{i=1}^{S}\pi_i\,(1-\pi_i)^t}{\sum_{i=1}^{S}\pi_i}$
(b) Rarefaction estimator for $t<T$		
$\hat{S}(t)=S_{obs}-\sum_{1\le Y_i\le T-t}\dfrac{\binom{T-Y_i}{t}}{\binom{T}{t}}$	$\widehat{PD}(t)=PD_{obs}-\sum_{1\le Y_i\le T-t}L_i\dfrac{\binom{T-Y_i}{t}}{\binom{T}{t}}$	$\hat{C}(t)=1-\sum_{1\le Y_i\le T-t}\dfrac{Y_i}{U}\dfrac{\binom{T-Y_i}{t}}{\binom{T-1}{t}}$
(c) Reference sample of size T		
$\hat{S}(T)=S_{obs}$	$\widehat{PD}(T)=PD_{obs}$	$\hat{C}(T)=1-\dfrac{Q_1}{U}\left[\dfrac{(T-1)Q_1}{(T-1)Q_1+2Q_2}\right]$
(d) Extrapolation estimator for sample size $T+t^*$		
$\hat{S}(T+t^*)=S_{obs}+\hat{Q}_0\left[1-\left(1-\dfrac{Q_1}{T\hat{Q}_0+Q_1}\right)^{t^*}\right]$	$\widehat{PD}(T+t^*)=PD_{obs}+\hat{R}_0\left[1-\left(1-\dfrac{R_1}{T\hat{R}_0+R_1}\right)^{t^*}\right]$	$\hat{C}(T+t^*)=1-\dfrac{Q_1}{U}\left[\dfrac{(T-1)Q_1}{(T-1)Q_1+2Q_2}\right]^{t^*+1}$

Notes: $U=\sum_{Y_i>0}Y_i=\sum_{j=1}^{T}jQ_j$ denotes the total number of incidences in T sampling units; $\hat{Q}_0$ and $\hat{R}_0$ denote the estimated number of undetected species richness in Eq. (2b) and undetected PD in Eq. (11c).

3.1. Sample-size-based rarefaction and extrapolation

Following Colwell et al. (2012), we refer to the observed sample of T sampling units as a *reference sample*. Let $S(t)$ be the expected number of species in a hypothetical sample of t sampling units, randomly selected from the sampling units that represent the assemblage. If we knew the true species detection probabilities $(\pi_1, \pi_2, \ldots, \pi_S)$ of the S species in each sampling unit, we could compute the following expected value:

$$S(t) = S - \sum_{i=1}^{S} (1-\pi_i)^t, \quad t = 1, 2, \ldots \tag{7a}$$

The plot of $S(t)$ with respect to the number of sampling units t is the sampling-unit-based species accumulation curve. Note that the true species richness represents the "asymptote" of the curve, i.e. $S = S(\infty)$. The rarefaction (interpolation) part estimates the expected species richness for a smaller number of sampling units $t < T$. On the basis of a reference sample of T sampling units, an unbiased estimator $\hat{S}(t)$ for $S(t)$, $t < T$, is

$$\hat{S}(t) = S_{obs} - \sum_{1 \le Y_i \le T-t} \binom{T-Y_i}{t} \Big/ \binom{T}{t}, \quad t < T. \tag{7b}$$

This analytic formula was first derived by Shinozaki (1963) and rediscovered multiple times (Chiarucci et al., 2008).

The extrapolation is to estimate the expected number of species $S(T+t^*)$ in a hypothetical sample of $T+t^*$ sampling units ($t^* > 0$) from the assemblage. Rewrite

$$\begin{aligned} S(T+t^*) &= \sum_{i=1}^{S} [1-(1-\pi_i)^{T+t^*}] \\ &= \sum_{i=1}^{S} [1-(1-\pi_i)^{T}] + \sum_{i=1}^{S} [1-(1-\pi_i)^{t^*}](1-\pi_i)^T \\ &= E\ (S_{obs}) + E\left[\sum_{i=1}^{S} [1-(1-\pi_i)^{t^*}] I(Y_i = 0)\right]. \end{aligned}$$

The first term in the above formula represents the observed species richness. For the second term, we can apply the Good-Turing incidence frequency formula (Section 2.3) by assuming that all previously undetected species have mean detection probability ϕ_0. Then for the second term, we have

$$\sum_{i=1}^{S} [1-(1-\pi_i)^{t^*}] I(Y_i = 0) \approx Q_0 [1-(1-\phi_0)^{t^*}].$$

Based on Eq. (5d), we have the extrapolated species richness for a sample of size $T+t^*$:

$$\hat{S}(T+t^*) = S_{obs} + \hat{Q}_0 \left[1 - \left(1 - \frac{Q_1}{T\hat{Q}_0 + Q_1}\right)^{t^*}\right], \quad t^* \geq 0. \tag{7c}$$

Colwell et al. (2012) linked rarefaction and extrapolation to form an integrated smooth curve. The integrated sample-size-based sampling curve includes a rarefaction part (which plots $\hat{S}(t)$ as a function of $t < T$), and an extrapolation part (which plots $\hat{S}(T+t^*)$ as a function of $T+t^*$), joining smoothly at the reference point (T, S_{obs}). The confidence intervals based on the bootstrap method (Section 3.3) also join smoothly.

For a short-range prediction (e.g. t^* is much less than T), the extrapolation formula is independent of the choice of $\hat{Q}_0$ as indicated by the approximation formula $\hat{S}(T+t^*) \approx S_{obs} + (Q_1/T)t^*$. This implies that the extrapolation formula in Eq. (7c) is very robust and reliable even though the species richness estimator is subject to bias. Previous experiences by Colwell et al. (2012) suggested that the prediction size can be extrapolated at most to double the observed sample size.

3.2. Coverage-based rarefaction and extrapolation

Turing and Good developed the very important concept of "sample coverage" to characterize the sample completeness of an observed set of individual-based abundance data. Their concept was extended by Chao et al. (1992) to capture-recapture data. For multiple incidence data, the *sample coverage* of a reference sample of T sampling units is defined as

$$C \equiv C(T) = \frac{\sum_{i=1}^{S} \pi_i I(Y_i > 0)}{\sum_{i=1}^{S} \pi_i} = 1 - \frac{\sum_{i=1}^{S} \pi_i I(Y_i = 0)}{\sum_{i=1}^{S} \pi_i},$$

which represents the fraction of the total incidence probabilities in the assemblage (including undetected species) that is represented by species detected in the reference sample. Note that under the binomial model (Eq. 1b), an unbiased estimator for the denominator in $C(T)$ is U/T, where $U = \sum_{k=1}^{T} kQ_k = \sum_{i=1}^{S} Y_i$ denotes the total number of incidences in the reference sample. For the numerator, we can apply the Good-Turing incidence frequency formula (Section 2.3) by assuming that all uniques in the sample have approximately the same detection probabilities, ϕ_1. Then we can write

$$\begin{aligned} E\left[\sum_{i=1}^{S} \pi_i I(Y_i = 0)\right] &= \sum_{i=1}^{S} \pi_i (1-\pi_i)^T \\ &= \frac{1}{T} E\left[\sum_{i=1}^{S} (1-\pi_i) I(Y_i = 1)\right] \approx \frac{E(Q_1)}{T}(1-\phi_1). \end{aligned}$$

Applying the Good-Turing formula $\hat{\phi}_1 = 2Q_2/[2Q_2 + (T-1)Q_1]$ (Eq. 5d), we obtain a very accurate estimator of the sample coverage for the reference sample size, if $Q_2 > 0$:

$$\hat{C}(T) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1 + 2Q_2}\right]. \tag{7d}$$

If $Q_2 = 0$, a modified formula based on Chao et al. (2014b, Appendix G) is:

$$\hat{C}(T) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)(Q_1-1)}{(T-1)(Q_1-1) + 2}\right]. \tag{7e}$$

In addition to the reference sample, we also need to consider the estimation of the expected sample coverage, $E[C(t)]$, for any hypothetical sample of t sampling units, $t = 1, 2, \ldots$. This expected sample coverage is a function of t as given below:

$$E[C(t)] = 1 - \frac{\sum_{i=1}^{S} \pi_i (1-\pi_i)^t}{\sum_{i=1}^{S} \pi_i}, \quad t \geq 1. \tag{7f}$$

For a rarefied sample ($t < T$), an unbiased estimator exists for the denominator and numerator in Eq. (7f), respectively, but their ratio $\hat{C}(t)$, given below, is only a nearly unbiased estimator of $E[C(t)]$:

$$\hat{C}(t) = 1 - \sum_{1 \leq Y_i \leq T-t} \frac{Y_i}{U} \frac{\binom{T-Y_i}{t}}{\binom{T-1}{t}}, \quad t < T.$$

An estimator for the expected coverage of an extrapolated sample with $T + t^*$ sampling units if $Q_2 > 0$ is

$$\hat{C}(T+t^*) = 1 - \frac{Q_1}{U}\left[\frac{(T-1)Q_1}{(T-1)Q_1 + 2Q_2}\right]^{t^*+1}. \tag{7g}$$

The above estimator is based on the following approximation formula:

$$E[C(T+t^*)] = 1 - \frac{\sum_{i=1}^{S} \pi_i (1-\pi_i)^{T+t^*}}{\sum_{i=1}^{S} \pi_i} \approx 1 - \frac{E[\sum_{i=1}^{S} (1-\pi_i)^{t^*+1} I(Y_i = 1)]}{T \sum_{i=1}^{S} \pi_i},$$

$$\approx 1 - \frac{[E(Q_1)](1-\phi_1)^{t^*+1}}{T \sum_{i=1}^{S} \pi_i}.$$

Replacing $\sum_{i=1}^{S} \pi_i$ and ϕ_1 with their respective estimators, U/T and $\hat{\phi}_1 = 2Q_2/[2Q_2 + (T-1)Q_1]$, we obtain Eq. (7g). If $Q_2 = 0$, a similar modification as in Eq. (7e) can be applied. Note that when $t^* = 0$, Eq. (7g) reduces to the sample coverage estimator for the reference sample. The coverage-based sampling curve includes a rarefaction part (which plots $\hat{S}(t)$ as a function of $\hat{C}(t)$), and an extrapolation part (which plots $\hat{S}(T+t^*)$ as a function of $\hat{C}(T+t^*)$), joining smoothly at the reference sample point

$(\hat{C}(T), S_{obs})$. The confidence intervals based on the bootstrap method (Section 3.3) also join smoothly. To equalize coverage among multiple, independent reference samples, their coverage-based curves can be extended to the coverage of the maximum size used in the corresponding sample-size-based sampling curve.

The sample-size-based approach plots the estimated species richness as a function of sample size, whereas the corresponding coverage-based approach plots the same richness estimate with respect to sample coverage. Therefore, the two types of sampling curves can be bridged by a sample completeness curve, which shows how the sample coverage estimate varies with sample size and also provides an estimate of the sample size needed to achieve a fixed degree of completeness. The two types of sampling curves along with the associated sample completeness curve are illustrated in Section 7 through an example. There, we also illustrate the use of the online software iNEXT (iNterpolation/EXTrapolation) to compute and plot the integrated sampling curves for incidence data. These methods allow researchers to efficiently use all available data to make more robust and more detailed inferences about species richness of the sampled assemblages, and also to make objective comparisons of species richness across assemblages.

3.3. Bootstrap method to obtain variance estimator and confidence intervals

The interpolated and extrapolated estimators are complicated functions of incidence data. Thus, it is not possible to derive analytic variance estimators. A bootstrap procedure can be applied to approximate the variance of any estimator based on incidence data. The estimated variance estimator can be subsequently used to construct a confidence interval of the expected species richness. Here we use the rarefied estimator $\hat{S}(t)$ given in Eq. (7b) as an example. Parallel steps can be formulated for any extrapolated estimator, coverage estimators, and for Chao2-type estimators.

First, we construct the *bootstrap assemblage*, which aims to mimic the true entire assemblage. Given a reference sample of size T and species sample incidence frequencies $(Y_1, Y_2, \ldots, Y_S)$, let $\hat{Q}_0$ be the Chao2-type estimator of the number of undetected species. Since the number of species in the bootstrap assemblage must be an integer, we define $\hat{Q}_0^*$ as the smallest integer that is greater than or equal to $\hat{Q}_0$. Thus, there are $S_{obs} + \hat{Q}_0^*$ species in the bootstrap assemblage.

Next we determine the detection probabilities in any sampling unit for the species in the bootstrap assemblage. Given that the ith species is detected in $Y_i > 0$ sampling units (there are S_{obs} of such species), the sample detection probability Y_i/T of an observed species ($Y_i > 0$), on average, overestimates the true detection probability π_i. This overestimation is due to the following conditional expectation:

$$E\left(\frac{Y_i}{T}\middle| Y_i > 0\right) = \frac{\pi_i}{1-(1-\pi_i)^T} > \pi_i.$$

The above conditional expectation leads to

$$\pi_i = E\left(\frac{Y_i}{T}\middle| Y_i > 0\right)[1-(1-\pi_i)^T].$$

If we replace the expected value in the above equation by the observed data, then we have the following approximation:

$$\pi_i \approx \frac{Y_i}{T}[1-(1-\pi_i)^T]. \tag{7h}$$

For any given $Y_i > 1$, one can numerically solve the above equation for π_i; but for $Y_i = 1$ (singletons, the most important count in our analysis), the only solution is $\pi_i = 0$, which is not reasonable. Therefore, Chao et al. (2014b, Appendix G) recommended the following analytic approach. Note that Eq. (7h) reveals that the approximate adjustment factor for the sample detection probability Y_i/T would be $[1-(1-\pi_i)^T]$. However, the adjustment factor $[1-(1-\pi_i)^T]$ cannot be estimated simply by substituting the sample detection probability for π_i, because the sample detection probability does not estimate π_i well for rare species. Chao et al. (2014b) suggested a more flexible adjustment factor, $[1-\tau(1-Y_i/T)^T]$. Applying this factor, we obtain that the species incidence probabilities for the S_{obs} observed species in the bootstrap assemblage can be estimated by

$$\hat{\pi}_i = \frac{Y_i}{T}\left[1-\hat{\tau}\left(1-\frac{Y_i}{T}\right)^T\right], \quad Y_i > 0, \tag{8a}$$

where $\hat{\tau}$ can be obtained from the sample coverage estimate:

$$\hat{C}(T) \times \frac{U}{T} = \sum_i \hat{\pi}_i I(Y_i > 0) = \sum_{Y_i > 0} \frac{Y_i}{T}\left[1-\hat{\tau}\left(1-\frac{Y_i}{T}\right)^T\right],$$

Then we can solve for $\hat{\tau}$:

$$\hat{\tau} = \frac{\frac{U}{T}[1-\hat{C}(T)]}{\sum\limits_{Y_i \geq 1} \frac{Y_i}{T}\left(1-\frac{Y_i}{T}\right)^T} = \frac{[1-\hat{C}(T)]}{\sum\limits_{Y_i \geq 1} \frac{Y_i}{U}\left(1-\frac{Y_i}{T}\right)^T}. \tag{8b}$$

We assume that each of the remaining $\hat{Q}_0^*$ species in the bootstrap assemblage (i. e. those species that were not detected in any sampling unit but exist in the bootstrap assemblage) has a common detection probability of $(U/T)[1-\hat{C}(T)]/\hat{Q}_0^*$. This assumption may seem restrictive, but the effect on the resulting variance estimator is limited, based on our extensive simulations.

After the bootstrap assemblage is determined, a random sample of T sampling units is generated from the assemblage, and a bootstrap estimate $\hat{S}(t)$ is calculated for the generated sample. The procedure is repeated B times to obtain B bootstrap estimates ($B = 200$ is suggested). The bootstrap variance estimator $\hat{S}(t)$ is the sample variance of these B estimates. The resulting bootstrap *s.e.* of $\hat{S}(t)$ is then used to construct a 95% confidence interval $\hat{S}(t) \pm 1.96$ *s.e.* $[\hat{S}(t)]$ for the expected species richness in a sample of size t. Similar procedures can be used to derive variance estimators for any other estimator and its associated confidence intervals.

4. Species richness estimation under sampling without replacement

Chao's original inequality was developed under the binomial (Eq. 1b) model, which assumes that sampling units are taken with replacement. When sampling is done without replacement, e.g. quadrats or time periods that are not repeatedly selected/surveyed, or mobile species are collected by lethal sampling methods, Chao's inequality and the Chao2 estimator require modification, unless the sampling fraction is small. For simplicity, we assume quadrat sampling in the following derivation, but the term "quadrat," here, may refer to any sampling unit that is not sampled with replacement, such as a trap, net, team, observer, occasion, transect line, or fixed period of time in other sampling protocols. Suppose that the region under investigation consists of T^* disjoint, equal-area quadrats, and a sample of T quadrats is randomly selected. Then each quadrat is surveyed, and species detection/non-detection data are recorded for each of these T quadrats.

The model assumes that species i can be detected in only U_i quadrats (U_i is unknown). We restrict our analysis to the case $U_i > 1$. (For any species with $U_i = 0$, there is no chance to detect this species in any sample, so it should be excluded from the estimating target.) In the other $T^* - U_i$ quadrats, species i is either absent or it is present but cannot be detected. Because U_i may vary independently among species, our model holds even if species are spatially aggregated, associated, or dissociated in the study area.

Assume that detection/non-detection of all species for each of the T quadrats is recorded to form a species-by-quadrat incidence matrix. Using the same notation as in Section 2, we let Y_i (sample incidence frequency) be the number of quadrats in which the ith species is observed in the sample, $i = 1, 2, \ldots, S$. Under sampling without replacement, the sample frequencies $(Y_1, Y_2, \ldots, Y_S)$ given $U_i = u_i$, follow a product-hypergeometric distribution:

$$P(Y_i = y_i, i = 1, 2, \ldots, S) = \prod_{i=1}^{S} \left\{ \binom{u_i}{y_i} \binom{T^* - u_i}{T - y_i} \Big/ \binom{T^*}{T} \right\}, \quad 1 \le u_i \le T^*. \qquad (9a)$$

That is, $(Y_1, Y_2, \ldots, Y_S)$ are independent but non-identically distributed random variables, each of which follows a hypergeometric distribution. If the sampling fraction is relatively small (i.e. $T^* \gg T$), then equation (9a) approaches the product binomial distribution:

$$P(Y_i = y_i, i = 1, 2, \ldots, S) \to \prod_{i=1}^{S} \left\{ \binom{T}{y_i} \left(\frac{u_i}{T^*}\right)^{y_i} \left(1 - \frac{u_i}{T^*}\right)^{T-y_i} \right\}.$$

This is a model for sampling with replacement with incidence probabilities $\pi_i = u_i/T^*$. The above approximation shows that, if there are many quadrats, and only a small number of the quadrats are sampled, then the inferences for the two types of sampling schemes differ little. Based on the general model (9a), the marginal distribution for each species' frequency is a hypergeometric distribution. The expected value of the frequency counts is

$$E(Q_k) = \sum_{i=1}^{S} P(Y_i = k) = \sum_{i=1}^{S} \frac{\binom{u_i}{k}\binom{T^* - u_i}{T - k}}{\binom{T^*}{T}}. \tag{9b}$$

In particular, we have

$$E(Q_0) = \sum_{i=1}^{S} \frac{\binom{T^* - u_i}{T}}{\binom{T^*}{T}},$$

$$E(Q_1) = \sum_{i=1}^{S} \frac{\binom{u_i}{1}\binom{T^* - u_i}{T - 1}}{\binom{T^*}{T}} = \sum_{i=1}^{S} \frac{T u_i}{T^* - u_i - T + 1} \frac{\binom{T^* - u_i}{T}}{\binom{T^*}{T}}$$

$$E(Q_2) = \sum_{i=1}^{S} \frac{\binom{u_i}{2}\binom{T^* - u_i}{T - 2}}{\binom{T^*}{T}} = \sum_{i=1}^{S} \frac{T(T-1)u_i(u_i - 1)}{2(T^* - u_i - T + 1)(T^* - u_i - T + 2)} \frac{\binom{T^* - u_i}{T}}{\binom{T^*}{T}}$$

The Cauchy-Schwarz inequality leads to

$$\left\{ \sum_{i=1}^{S} \frac{\binom{T^* - u_i}{T}}{\binom{T^*}{T}} \right\} \left\{ \sum_{i=1}^{S} \left(\frac{T u_i}{T^* - u_i - T + 1} \right)^2 \frac{\binom{T^* - u_i}{T}}{\binom{T^*}{T}} \right\} \geq \left\{ \sum_{i=1}^{S} \frac{T u_i}{T^* - u_i - T + 1} \frac{\binom{T^* - u_i}{T}}{\binom{T^*}{T}} \right\}^2,$$

The right side in the above inequality is $\{E(Q_1)\}^2$, and the first sum on the left side is $E(Q_0)$. For the second sum, we rewrite

$$\left(\frac{Tu_i}{T^*-u_i-T+1}\right)^2 = \frac{T}{T-1}\left(\frac{T(T-1)u_i(u_i-1)}{(T^*-u_i-T+1)^2}\right)+\frac{T^2u_i}{(T^*-u_i-T+1)^2}.$$

Thus the second sum becomes

$$\left\{\sum_{i=1}^{S}\left(\frac{Tu_i}{T^*-u_i-T+1}\right)^2\frac{\binom{T^*-u_i}{T}}{\binom{T^*}{T}}\right\}$$

$$\approx \frac{2T}{T-1}E(Q_2)+\sum_{i=1}^{S}\left[\frac{T}{T^*-u_i-T+1}\right]\frac{Tu_i}{T^*-u_i-T+1}\frac{\binom{T^*-u_i}{T}}{\binom{T^*}{T}}.$$

The contribution of species with large u_i (frequent species) to any term involved in the above Cauchy-Schwarz inequality is almost negligible. For infrequent species (with u_i much less than T^*), we have

$$\frac{T}{T^*-u_i-T+1} = \frac{T/T^*}{(T^*-u_i-T+1)/T^*} \approx \frac{T/T^*}{1-(T/T^*)} = \frac{q}{1-q},$$

where $q = T/T^*$ denotes the sampling fraction. We then obtain the following approximate inequality

$$\{E(Q_0)\}\left(\frac{T}{T-1}2E(Q_2)+\frac{q}{1-q}E(Q_1)\right) \geq \{E(Q_1)\}^2,$$

which is equivalent to

$$E(Q_0) \geq \frac{\{E(Q_1)\}^2}{\frac{T}{T-1}2E(Q_2)+\frac{q}{1-q}E(Q_1)}.$$

Replacing the expected value by the observed frequencies, we thus obtain the following lower bound for the true species richness.

$$\hat{S}_{wor2} = S_{obs} + \frac{Q_1^2}{2wQ_2+rQ_1}, \tag{9c}$$

where $w = T/(T-1)$ and $r = q/(1-q)$, and the subscript "*wor*" refers to "without replacement". When the sample fraction q approaches zero, then r approaches zero, and our lower bound approaches the Chao2 estimator. On the other hand, when q approaches 1, $r = q/(1-q)$ approaches infinity and our lower bound reduces to the number of observed species, which is the true parameter for complete sampling.

An approximate variance formula for $\hat{S}_{wor2}$ can be obtained by using an asymptotic approach based on the hypergeometric distribution. The resulting variance estimator is:

$$\widehat{\text{var}}(\hat{S}_{wor2}) = \hat{Q}_0 + \frac{(2wQ_2\hat{Q}_0^2 + Q_1^2\hat{Q}_0)^2}{Q_1^5} + 4w^2Q_2\left(\frac{\hat{Q}_0}{Q_1}\right)^4,$$

where $\hat{Q}_0 = \hat{S}_{wor2} - S_{obs}$ denotes the estimator of the undetected species in the sample. When $\hat{S}_{wor2}$ is used as an estimator of species richness, a confidence interval of S can be constructed by a log-transformation (Eq. 3c), so that the lower bound is always greater than the number of observed species.

5. Shared species richness estimation

We now extend the one-assemblage model formulation and data framework to two assemblages (I and II), which can differ not only in their species richness, but also in their species composition. Suppose that there are S species in the *pooled* assemblage. Assume that T_1 sampling units (Sample I) are randomly taken from Assemblage I, and T_2 sampling units (Sample II) are taken from Assemblage II. In each sampling unit, only species detection/non-detection data are recorded. The two sets of probabilities $(\pi_{11}, \pi_{21}, \ldots, \pi_{S1})$ and $(\pi_{12}, \pi_{22}, \ldots, \pi_{S2})$ in the incidence case represent species detection probabilities in any sampling unit from Assemblages I and II, respectively, $\pi_{i1}, \pi_{i2} \geq 0$, $i = 1, 2, \ldots, S$. Let the true number of shared species between the two assemblages be S_{12}. Without loss of generality, we assume that the first S_{12}species in the pooled assemblage are these shared species.

Let Y_{i1} and Y_{i2} denote the number of sampling units in which the ith species is detected in Samples I and II, respectively. For any two non-negative integers r and v, define

$$Q_{rv} = \sum\nolimits_{i=1}^{S_{12}} I(Y_{i1} = r, Y_{i2} = v), \quad r,\ v = 0, 1, 2, \ldots$$

That is, Q_{rv} denotes the number of *shared* species that are detected in r sampling units in Sample I and v sampling units in Sample II. In particular, Q_{11} denotes the number of shared species that are uniques in both samples, and Q_{00} denotes the number of shared species that are present in both samples, but detected in neither. Also, let Q_{r+} denote the number of shared species that are detected in r sampling units in Sample I and that are detected in at least one sampling unit (using a "+" sign to replace the index v) in Sample II, with a similar symmetric definition for Q_{+v}. Thus, Q_{++} becomes the total number of observed species shared between the two samples. Mathematically, we have the following expressions:

$$Q_{r+} = \sum\nolimits_{i=1}^{S_{12}} I(Y_{i1} = r,\ Y_{i2} \geq 1) = \sum\nolimits_{v>0} Q_{rv}, \quad r = 0, 1, 2, \ldots$$

$$Q_{+v} = \sum\nolimits_{i=1}^{S_{12}} I(Y_{i1} \geq 1,\, Y_{i2} = v) = \sum\nolimits_{r>0} Q_{rv}, \quad v = 0, 1, 2, \ldots$$

Here, Q_{+0} denotes the number of *shared* species that are detected in Sample I but not detected in Sample II, and a similar interpretation for Q_{0+}.

Since $S_{12} = Q_{++} + Q_{+0} + Q_{0+} + Q_{00}$ but only Q_{++} is observable, our approach is to find a lower bound for each of the expected values of the other three terms, i.e. $E(Q_{+0})$, $E(Q_{0+})$ and $E(Q_{00})$. Assuming the binomial models (Eq. 1b) for species incidence frequencies for each of the two independent sets of frequencies, we have

$$E(Q_{00}) = \sum\nolimits_{i=1}^{S_{12}} (1-\pi_{i1})^{T_1}(1-\pi_{i2})^{T_2},$$

$$E(Q_{+0}) = \sum\nolimits_{i=1}^{S_{12}} [1-(1-\pi_{i1})^{T_1}](1-\pi_{i2})^{T_2},$$

$$E(Q_{0+}) = \sum\nolimits_{i=1}^{S_{12}} (1-\pi_{i1})^{T_1}[1-(1-\pi_{i2})^{T_2}].$$

We now derive a lower bound for each term as follows.

1. A lower bound for $E(Q_{+0})$: Since

$$E(Q_{+1}) = \sum\nolimits_{i=1}^{S_{12}} [1-(1-\pi_{i1})^{T_1}]\; T_2\; \pi_{i2}(1-\pi_{i2})^{T_2-1},$$

$$E(Q_{+2}) = \sum\nolimits_{i=1}^{S_{12}} [1-(1-\pi_{i1})^{T_1}]\; [T_2(T_2-1)/2]\pi_{i2}^2(1-\pi_{i2})^{T_2-2}.$$

The following Cauchy-Schwarz inequality

$$\left[\sum\nolimits_{i=1}^{S_{12}} [1-(1-\pi_{i1})^{T_1}](1-\pi_{i2})^{T_2}\right]\left[\sum\nolimits_{i=1}^{S_{12}} [1-(1-\pi_{i1})^{T_1}]\; \pi_{i2}^2(1-\pi_{i2})^{T_2-2}\right]$$
$$\geq \left[\sum\nolimits_{i=1}^{S_{12}} [1-(1-\pi_{i1})^{T_1}]\; \pi_{i2}(1-\pi_{i2})^{T_2-1}\right]^2$$

leads to a lower bound

$$E(Q_{+0}) \geq \frac{(T_2-1)}{T_2}\frac{[E(Q_{+1})]^2}{2E(Q_{+2})}. \tag{10a}$$

2. Similarly, a lower bound for $E(Q_{0+})$ is

$$E(Q_{0+}) \geq \frac{(T_1-1)}{T_1}\frac{[E(Q_{1+})]^2}{2E(Q_{2+})}. \tag{10b}$$

3. A lower bound for $E(Q_{00})$ is obtained by noting

$$E(Q_{11}) = \sum\nolimits_{i=1}^{S_{12}} T_1\; \pi_{i1}(1-\pi_{i1})^{T_1-1}\; T_2\; \pi_{i2}(1-\pi_{i2})^{T_2-1},$$

$$E(Q_{22}) = \sum\nolimits_{i=1}^{S_{12}} [T_1(T_1-1)/2]\pi_{i1}^2(1-\pi_{i1})^{T_1-2}[T_2(T_2-1)/2]\pi_{i2}^2(1-\pi_{i2})^{T_2-2}.$$

Again, a similar Cauchy-Schwarz inequality

$$\left[\sum\nolimits_{i=1}^{S_{12}} (1-\pi_{i1})^{T_1}(1-\pi_{i2})^{T_2}\right] \left[\sum\nolimits_{i=1}^{S_{12}} \pi_{i1}^2(1-\pi_{i1})^{T_1-2}\pi_{i2}^2(1-\pi_{i2})^{T_2-2}\right]$$
$$\geq \left[\sum\nolimits_{i=1}^{S_{12}} \pi_{i1}(1-\pi_{i1})^{T_1-1}\pi_{i2}(1-\pi_{i2})^{T_2-1}\right]^2$$

gives

$$E(Q_{00}) \geq \frac{(T_1-1)}{T_1}\frac{(T_2-1)}{T_2}\frac{[E(Q_{11})]^2}{4E(Q_{22})}. \tag{10c}$$

Combining the above three lower bounds and letting $K_i = (T_i - 1)/T_i$, we thus have a lower bound for the shared species richness:

$$\hat{S}_{12} = Q_{++} + K_2\frac{Q_{+1}^2}{2Q_{+2}} + K_1\frac{Q_{1+}^2}{2Q_{2+}} + K_1K_2\frac{Q_{11}^2}{4Q_{22}}. \tag{10d}$$

The above estimator is referred to as the *Chao*2-*shared* estimator because it can be regarded as an extension of the single-assemblage Chao2 estimator (Eq. 2b) to the case of two assemblages. A bias-corrected estimator to avoid zero divisor is

$$\tilde{S}_{12} = Q_{++} + K_2\frac{Q_{+1}(Q_{+1}-1)}{2(Q_{+2}+1)} + K_1\frac{Q_{1+}(Q_{1+}-1)}{2(Q_{2+}+1)} + K_1K_2\frac{Q_{11}(Q_{11}-1)}{4(Q_{22}+1)}. \tag{10e}$$

Note that only observed, shared species are involved in the formulas (10a) to (10e), thus observed non-shared species play no role in our estimation, although any species observed in one Sample but not in the other could actually be a shared species. Because the proposed estimator can be regarded as a function of the statistics $(Q_{++}, Q_{11}, Q_{22}, Q_{1+}, Q_{2+}, Q_{+1}, Q_{+2})$, we obtain a variance estimator by using a standard asymptotic approach under a multinomial distribution. Then the estimated variance can be used to construct a confidence interval for the true parameter using a log-transformation (Chao, 1987).

The above approach has an obvious extension to the case of more than two assemblages. For example, in the case of three assemblages, a "shared" species is defined as that the species belongs to all three assemblages. Assume that there are S_{123} species shared by all three assemblages (I, II and III), and a random sample of sampling units is taken from each of the three assemblages. The three samples are called Samples I, II and III with sizes T_1, T_2 and T_3 respectively. Then

$$S_{123} = Q_{+++} + Q_{++0} + Q_{+0+} + Q_{0++} + Q_{00+} + Q_{0+0} + Q_{+00} + Q_{000},$$

where Q_{+++} denotes the observed shared species richness in the three samples, Q_{++0} denotes the number of shared species that are observed in Samples I, II but not observed in Sample III, Q_{000} denotes the number of shared species that are not detected in any of the three samples, and a similar interpretation for other terms in the above formula. Parallel derivations (with self-explanatory notation) lead to a lower bound for S_{123} as follows:

$$\begin{aligned}\hat{S}_{123} = Q_{+++} + K_3\frac{Q_{++1}^2}{2Q_{++2}} + K_2\frac{Q_{+1+}^2}{2Q_{+2+}} + K_1\frac{Q_{1++}^2}{2Q_{2++}} \\ + K_1K_2\frac{Q_{11+}^2}{4Q_{22+}} + K_1K_3\frac{Q_{1+1}^2}{4Q_{2+2}} + K_2K_3\frac{Q_{+11}^2}{4Q_{+22}} + K_1K_2K_3\frac{Q_{111}^2}{8Q_{222}}.\end{aligned}$$

We can formulate a bias-corrected version to avoid zero divisor in the same manner as that given in Eq. (10e). An estimated variance can be obtained by an asymptotic method.

6. Phylogenetic richness estimation

6.1. Framework

In traditional measures of species diversity, all species (or taxa at some other rank) are considered to be equally distinct from one another. However, in an evolutionary context, species differences can be based directly on their evolutionary relationships, either in the form of taxonomic classification or well-supported phylogenetic trees. Species that are closely related are generally less distinct in important ecological characteristics than are distantly-related species. A wide range of phylogenetic diversity metrics and related (dis)similarity measures have been proposed in the literature. The most widely used phylogenetic metric is Faith's (1992) *PD* (phylogenetic diversity), which is defined as the sum of the branch lengths of a phylogenetic tree connecting all species in the focal assemblage.

Chao et al. (2010, 2015) proposed a class of abundance-sensitive phylogenetic measures and showed that Faith's *PD* is a phylogenetic generalization of species richness. In other words, Faith's *PD* is a phylogenetic diversity of order zero in which species abundances are not considered. From this perspective, Faith's *PD* is a measure of *phylogenetic richness*. Throughout this paper, *PD* refers to Faith's (1992) *PD*. When some species that are present in an assemblage are not detected in a sample, the lineages/branches associated with these undetected species are also missing from the phylogenetic tree of the observed species. The undetected *PD* in an incomplete sample was not discussed until recent years (Cardoso et al., 2014; Chao et al., 2015).

Model formulation and *PD* estimation based on abundance data were developed in Chao et al. (2015). The corresponding framework for incidence data, introduced in their Appendix S7 and presented here, is a generalization of the framework for species richness. As discussed in Section 2.1, suppose, in the focal assemblage, that there are S species indexed by $1,2,\dots,S$, and T sampling units are surveyed from the assemblage. In each sampling unit, we assume that only incidence (detection or non-detection) of each species is recorded. For any sampling unit, assume that the ith species has its own unique incidence (or detection) probability π_i that is constant for any randomly selected sampling unit. We also assume that a rooted ultrametric or non-ultrametric phylogenetic tree of the S species (as tip nodes) can be constructed. Here we assume that all phylogenetic measures are computed from a fixed, basal reference point in the tree that is ancestral to all taxa considered in the study.

Assume that there are B branch segments in the corresponding tree, $B \geq S$, descendant to the given basal reference point. Let L_i denote the length of the ith branch. We expand the set of detection probabilities $(\pi_1, \pi_2, \dots, \pi_S)$ of the S species (as tip nodes) to a larger set of branch/node detection probabilities $\{\lambda_i,\ i = 1,2,\dots,B\}$ with $(\pi_1, \pi_2, \dots, \pi_S)$ as the first S elements. Here we define λ_i as the probability of detecting at least one species descended from branch i in a sampling unit, $i = 1,2,\dots,B$, and refer to λ_i as the incidence (or detection) probability of branch/node i. The true PD for the fixed reference point is expressed as $PD = \sum_{i=1}^{B} L_i$.

The species-by-sampling-unit incidence matrix $\{W_{ij};\ i = 1,2,\dots,S,\ j = 1,2,\dots,T\}$ and the species incidence frequencies $Y_i = \sum_{j=1}^{T} W_{ij}$ are defined exactly the same as those in Section 2.1. Here we expand the $S \times T$ incidence matrix $\{W_{ij};\ i = 1,2,\dots,S,\ j = 1,2,\dots,T\}$ to a larger $B \times T$ matrix $\{W_{ij}^*,\ i = 1,2,\dots,B,\ j = 1,2,\dots,T\}$ by specifying that $W_{ij}^* = 1$ if at least one species descended from branch i is detected in jth sampling unit, and $W_{ij}^* = 0$ otherwise. This specification also expands the set of the observed species incidence frequencies $\{Y_1, Y_2, \dots, Y_S\}$ to a larger set $\{Y_i^*,\ i = 1,2,\dots,B\}$, which consists of the row sums of the expanded incidence matrix $[W_{ij}^*]$. We refer to Y_i^* as the sample *branch/node incidence frequency* for branch/node i, $i = 1,2,\dots,B$. See Table 3 for a simple, hypothetical dataset for nine species in six sampling units, illustrating the expansion of the matrix $[W_{ij}]$ to $[W_{ij}^*]$.

Suppose that the incidence data for all the sampling units are independent. Then Y_i^*, $i = 1,2,\dots,B$, follows a binomial distribution:

$$P(Y_i^* = y_i) = \binom{T}{y_i} \lambda_i^{y_i} (1-\lambda_i)^{T-y_i}, \quad y_i = 0,1,2,\dots,T.$$

Define R_k as the sum of branch lengths for those branches with branch/node incidence frequency k, i.e.

$$R_k = \sum\nolimits_{i=1}^{B} L_i\, I(Y_i^* = k), \quad k = 0,1,\dots,T. \tag{11a}$$

Thus, R_0 represents the total length of branches that are not detected in the observed tree (i.e. not detected by the tree spanned by the observed species in the reference sample), and R_1 denotes the total branch length of the uniques in the branch incidence frequency set of the observed tree. A similar interpretation is valid for R_2. Let PD_{obs} denote the

Table 3: Species detection/non-detection data for the hypothetical tree in Figure 1. Species 4, 7, 8, and 9 (grey shaded area) are not observed in the sample; Node 14 (grey shaded area) is not observed in the tree spanned by the observed species.

Species/node/branch	Detection/non-detection in six sampling units (1 means detection; blank means non-detection)						Species/node/branch incidence frequency
	1	2	3	4	5	6	
1	1	1	1	1	1	1	$Y_1 = 6$
2		1					$Y_2 = 1$
3		1					$Y_3 = 1$
4							$Y_4 = 0$
5		1	1				$Y_5 = 2$
6				1			$Y_6 = 1$
7							$Y_7 = 0$
8							$Y_8 = 0$
9							$Y_9 = 0$
10		1					$Y_{10}^* = 1$
11	1	1	1	1	1	1	$Y_{11}^* = 6$
12		1	1				$Y_{12}^* = 2$
13				1			$Y_{13}^* = 1$
14							$Y_{14}^* = 0$
15				1			$Y_{15}^* = 1$

observed PD. Then we have $PD_{obs} = \sum_{i>0} R_i$ and $PD = PD_{obs} + R_0$. See Figure 1 for a hypothetical tree spanned by 9 species for an example.

6.2. Chao's inequality for PD

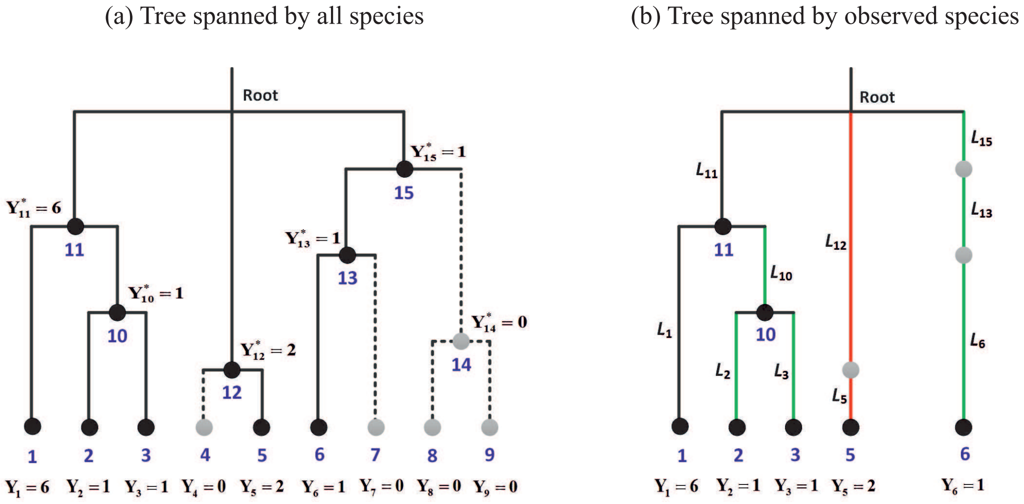
The undetected PD in the reference sample is R_0, which is unknown. However, $\{R_1, R_2, \ldots\}$ can be computed from the reference sample and the tree spanned by the observed species. Following the same approach that Chao et al. (2015) used for abundance data, we have the expected value of R_k:

$$E(R_k) = E\left[\sum_{i=1}^{B} L_i I(Y_i^* = k)\right] = \binom{T}{k} \sum_{i=1}^{B} L_i \lambda_i^k (1-\lambda_i)^{T-k}, \quad k = 0, 1, \ldots, T. \quad (11b)$$

In particular, we have

$$E(R_0) = \sum_{i=1}^{B} L_i (1-\lambda_i)^T,$$

$$E(R_1) = T \sum_{i=1}^{B} L_i \lambda_i (1-\lambda_i)^{T-1},$$


(a) Tree spanned by all species

(b) Tree spanned by observed species

Figure 1: *(a) A hypothetical tree spanned by 9 species (tip nodes) indexed by 1, 2, …, 9 in an assemblage. The ancestor of the entire assemblage is the "root" at the top, with time progressing towards the branch tips at the bottom. Here the root of the entire assemblage is selected as the reference point for illustration. Species detection/non-detection records in six sampling units are given in Table 3. A black dot means a node with species incidence frequency* > 0*; a grey dot means a node with frequency* $= 0$*. (b) A sub-tree spanned by the observed 5 species (1, 2, 3, 5 and 6). Species 4, 7, 8 and 9 are not detected in any of the six sampling units, so only a portion of the tree (solid branches in the left panel) is observed as shown in Panel (b). Black dots in Panel (b) are nodes in the observed tree; grey dots are not observed in the tree. The sample incidence frequency vector in 6 sampling units for 9 species is* $(Y_1, Y_2, \ldots, Y_9)$ *= (6, 1, 1, 0, 2, 1, 0, 0, 0); only non-zero frequencies represent observed species. The branch set B in the assemblage includes 15 branches (indexed from 1 to 15) with branch lengths* $(L_1, L_2, \ldots, L_{15})$ *and the corresponding 15 nodes. The corresponding node/branch incidence frequencies are* $(Y_1^*, Y_2^*, \ldots, Y_9^*, Y_{10}^*, Y_{11}^*, \ldots, Y_{15}^*)$ *= (6, 1, 1, 0, 2, 1, 0, 0, 0, 1, 6, 2, 1, 0, 1) with* $(Y_1, Y_2, \ldots, Y_9)$ *as the first 9 elements (see Table 3). The dotted branches in Panel (a) are not detected in the sample, and the total length of the undetected branches is* $R_0 = L_4 + L_7 + L_8 + L_9 + L_{14}$*. In Panel (b), the total length of those branches with* $Y_i^* = 1$ *(there are four uniques in the node/branch incidence frequency set of the observed tree) is* $R_1 = L_2 + L_3 + L_6 + L_{10} + L_{13} + L_{15}$ *(as shown by green lines in the observed tree in Panel (b)); the total length with* $Y_i^* = 2$ *(only one duplicate in the node/branch incidence frequency set of the observed tree) is* $R_2 = L_5 + L_{12}$ *(as shown by red lines in the observed tree in Panel (b)).*

$$E(R_2) = \frac{T(T-1)}{2} \sum_{i=1}^{B} L_i \lambda_i^2 (1-\lambda_i)^{T-2}.$$

The Cauchy-Schwarz inequality

$$\left[\sum_{i=1}^{B} L_i (1-\lambda_i)^T\right] \left[\sum_{i=1}^{B} L_i \lambda_i^2 (1-\lambda_i)^{T-1}\right] \geq \left[\sum_{i=1}^{B} L_i \lambda_i (1-\lambda_i)^{T-1}\right]^2$$

leads to the following inequality:

$$E(R_0) \geq \frac{(T-1)}{T} \frac{[E(R_1)]^2}{2E(R_2)}.$$

Thus, a direct estimator of the undetected *PD* would be $\frac{(T-1)}{T} \frac{R_1^2}{2R_2}$. However, when R_2 is relatively small, including the case of $R_2 = 0$, this estimator may yield an extremely large value and thus exhibit a large variance. To cope with such cases, Chao et al. (2015) and Hsieh and Chao (2017) proposed the following Chao2-*PD* estimator:

$$\widehat{PD}_{Chao2} = PD_{obs} + \hat{R}_0 = \begin{cases} PD_{obs} + \dfrac{(T-1)}{T} \dfrac{R_1^2}{2R_2}, & \text{if } R_2 > \dfrac{R_1 Q_2^*}{2Q_1^*}; \\[2ex] PD_{obs} + \dfrac{(T-1)}{T} \dfrac{R_1(Q_1^* - 1)}{2(Q_2^* + 1)}, & \text{if } R_2 \leq \dfrac{R_1 Q_2^*}{2Q_1^*}. \end{cases} \tag{11c}$$

where Q_1^* and Q_2^* denote, respectively, the number of nodes/branches with incidence frequency $= 1$ and frequency $= 2$ in the observed tree; see Figure 1 for an example.

As with the Chao2 estimator, this lower bound is a nearly unbiased point estimator if unique and undetected branches/nodes have approximately identical mean detection probabilities. A sufficient condition is that rare/infrequent node/branch detection probabilities are approximately homogeneous, while other nodes/branches can be highly heterogeneous. When the detection probabilities for rare nodes/branches are heterogeneous and the sample is not sufficiently large, negative bias exists. The variance of the Chao1-*PD* estimator can be obtained using Eqs. (3a) and (3b) with $\{Q_1, Q_2\}$ being replaced by $\{R_1, R_2\}$. The construction of the confidence interval for Faith *PD* based on the Chao1-*PD* estimator can be similarly obtained as that given in Eq. (3c).

Comparing the derivations for the above phylogenetic version of Chao's inequality with those in Section 2.3 for species richness, we see that all estimation steps are parallel and the analogy between the two estimation frameworks is transparent. The analogy was first proposed by Faith (1992). From Faith's perspective, each unit-length branch is regarded as a "feature" in phylogenetic diversity (like a "species" in species diversity). Chao et al. (2014a) subsequently referred to each unit-length branch segment as a *phylogenetic entity*. All entities are phylogenetically equally distinct, just as all species are assumed taxonomically equally distinct in computing simple species richness. Instead of species, for *PD* we are measuring the total number of phylogenetic entities, or equivalently, the total branch length (because each entity has length of unity). Based on this perspective, for incidence data the measures of branch lengths $\{R_k, k = 0, 1, \ldots, \}$ used to estimate *PD* play the same role as the frequency counts $\{Q_k, k = 0, 1, \ldots\}$ in estimating species richness. This analogy to counting up species means that most ecological indices defined at the species level can be converted to *PD* equivalents (by counting phylogenetic entities rather than species).

6.3. Rarefaction/extrapolation guided by the Chao2-PD estimator

Because of the analogy between counting up species richness and counting up branch lengths, all the species richness estimation tools for standardized samples in Section 3 can be directly extended to their phylogenetic equivalents, and similar sample-size-based and coverage-based rarefaction and extrapolation sampling curves can be constructed. Table 2 gives all the corresponding formulas; thus we omit all details except for the following two notes.

The theoretical formula for $PD(t)$, the expected PD when a set of t sampling units is taken from the assemblage, is a generalization of Eq. (7a):

$$PD(t) = \sum_{i=1}^{B} L_i[1 - (1 - \lambda_i)^t], \quad t = 1, 2, \ldots$$

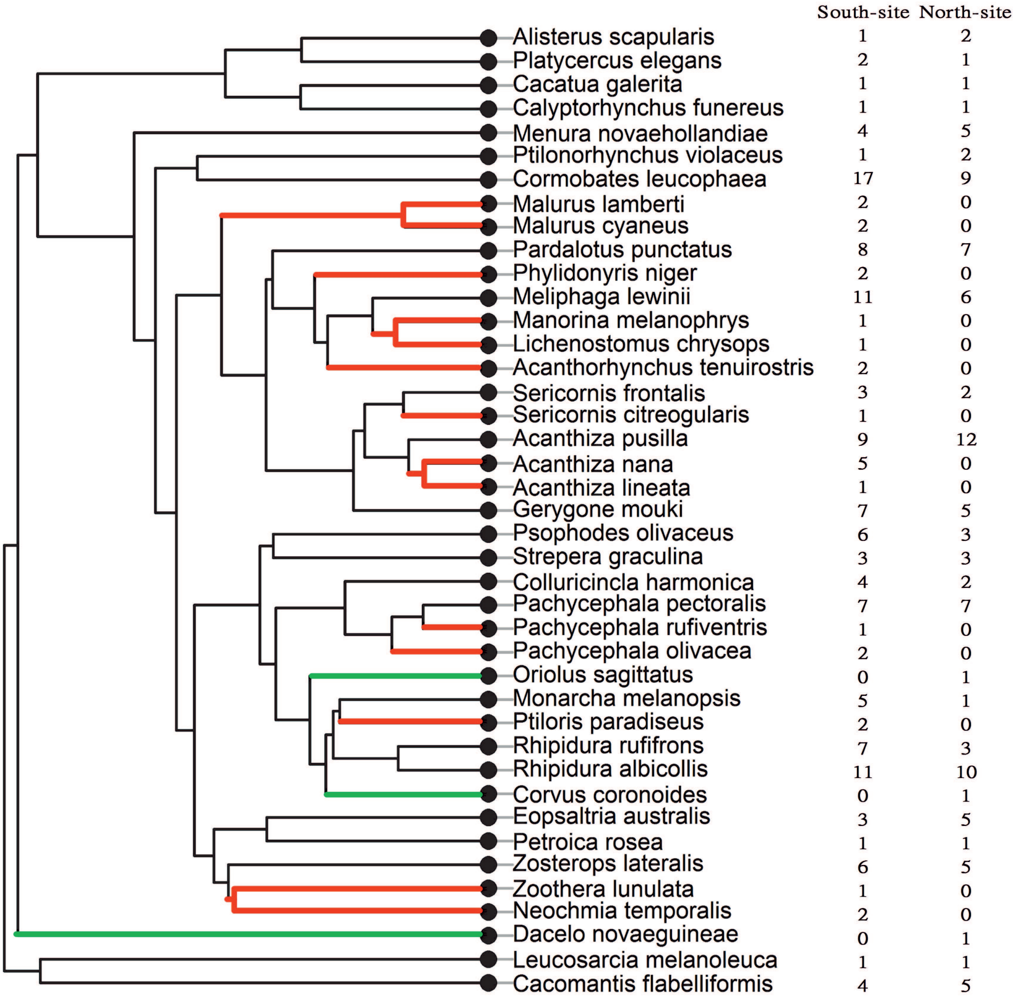
The plot of $PD(t)$ as a function of t is a non-decreasing function and is referred to as the *sampling-unit-based PD accumulation curve*. As sample size t tends to infinity, $PD(t)$ approaches the true PD. Thus the true PD represents the "asymptote" of the PD accumulation curve, i.e. the true $PD = PD(\infty)$. When there are no internal nodes, and all S lineages are equally distinct with branch lengths of unity (i.e. branch lengths are normalized to unity), the sampling-unit-based PD accumulation curve reduces to the species accumulation curve.

The bootstrap method to assess the variance and confidence interval associated with the PD estimator for rarefied and extrapolated samples is similar to that in Section 3.3, except that a "bootstrap tree" should be constructed in the resampling procedure. Recall that, in the bootstrap assemblage discussed in Section 3.3 for species richness, there are $S_{obs} + \hat{Q}_0^*$ species, where $\hat{Q}_0^*$ is the smallest integer that is greater than or equal to the estimated undetected species richness $\hat{Q}_0$ based on the Chao2 estimator in Eq. (2b). The PD bootstrap tree includes two portions: the known tree spanned by the observed species, and the undetected tree spanned by the remaining $\hat{Q}_0^*$ species in the bootstrap assemblage. The latter portion of tree is estimated by assuming that the undetected species in the bootstrap tree all diverged directly from the root of the observed tree with a constant branch length $\hat{\bar{L}}_{(0)}$, where $\hat{\bar{L}}_{(0)} = \hat{R}_0/\hat{Q}_0^*$, and $\hat{R}_0$ is the estimated undetected PD based on Eq. (11c). This augmented portion of tree may seem to be restrictive, but the effect on the resulting variance is limited; see Chao et al. (2015) for details.

7. Example

7.1. Data description (Figure 2, Appendices A and B)

A small empirical data set for birds observed in November 2012 in Australian Barrington Tops National Park is used for illustration. The original data were described in Chao et al. (2015). At each data sampling point, the abundance of each bird species observed



	South-site	North-site
Alisterus scapularis	1	2
Platycercus elegans	2	1
Cacatua galerita	1	1
Calyptorhynchus funereus	1	1
Menura novaehollandiae	4	5
Ptilonorhynchus violaceus	1	2
Cormobates leucophaea	17	9
Malurus lamberti	2	0
Malurus cyaneus	2	0
Pardalotus punctatus	8	7
Phylidonyris niger	2	0
Meliphaga lewinii	11	6
Manorina melanophrys	1	0
Lichenostomus chrysops	1	0
Acanthorhynchus tenuirostris	2	0
Sericornis frontalis	3	2
Sericornis citreogularis	1	0
Acanthiza pusilla	9	12
Acanthiza nana	5	0
Acanthiza lineata	1	0
Gerygone mouki	7	5
Psophodes olivaceus	6	3
Strepera graculina	3	3
Colluricincla harmonica	4	2
Pachycephala pectoralis	7	7
Pachycephala rufiventris	1	0
Pachycephala olivacea	2	0
Oriolus sagittatus	0	1
Monarcha melanopsis	5	1
Ptiloris paradiseus	2	0
Rhipidura rufifrons	7	3
Rhipidura albicollis	11	10
Corvus coronoides	0	1
Eopsaltria australis	3	5
Petroica rosea	1	1
Zosterops lateralis	6	5
Zoothera lunulata	1	0
Neochmia temporalis	2	0
Dacelo novaeguineae	0	1
Leucosarcia melanoleuca	1	1
Cacomantis flabelliformis	4	5

Figure 2: *The phylogenetic tree of 41 bird species and the sample species incidence frequencies for two sites (the North Site with 12 point-counts and the South Site with 17 point-counts) in Australian Barrington Tops National Park (Chao et al., 2015). The phylogenetic tree is a Maximum Clade Credibility tree from the Bayesian analysis of Jetz et al. (2012). Branch lengths are scaled to millions of years since divergence. The phylogenetic tree for the species observed in the North Site includes black branches and green branches. The phylogenetic tree for the species observed in the South Site includes black branches and red branches. (Black branches are shared by both sites; red and green branches are non-shared.) A zero-frequency in a site means that the species was not observed in that site. The age of the root (i.e. tree depth) is 82.9 millions of years.*

over a 30-minute period in a 50 m radius was recorded – called a *point-count* in ornithology. We treat each point-count as a sampling unit. There were 12 point counts conducted along the Barrington Tops Forest Road in the northern part of the national park. The corresponding records, shown in Appendix A, form the reference sample for the North Site. There were 17 point counts conducted along the Gloucester Tops

Road in the southern part of the Barrington Tops National Park; the raw detection/non-detection records (ignoring abundances) for the 17 point counts are listed in Appendix B. Those records form the reference sample for the South Site. Vegetation at both sites ranged from wet eucalypt forest to rainforest, with an average canopy cover of 80% for the North Site and 60% for the South Site. The sampling points comprising the North Site had an average elevation of 1078 m, while those of the South Site had an average elevation of 928 m. A total of 41 species were observed, for both sites combined, and all species incidence frequencies are shown in Figure 2 and in the last column of Appendices A and B. A phylogenetic tree of these species (Figure 2) was constructed from a Maximum Clade Credibility tree of the Bayesian analysis of Jetz et al. (2012). The age of the root for the phylogenetic tree spanned by the observed species is 82.9 million years (Myr). Chao et al. (2015) analyzed these data based on species abundance data. Here we focus on species incidence frequency data which can account for spatial heterogeneity in the data, whereas abundance-based approach often cannot.

7.2. Species richness and shared species richness estimation (Table 4)

In the North Site ($T = 12$ sampling units), the reference sample includes 102 incidences ($U = 102$) representing 27 observed species; in the South Site ($T = 17$ sampling units), the reference sample includes 148 incidences ($U = 148$) representing 38 observed species. The species incidence frequency counts (Q_1 to Q_T) for the two sites are summarized in Table 4. Based on Eq. (7d), the estimated sample coverage values for the North Site and the South Site are nearly identical at a level of 92% (specifically, 91.8% for the North Site and 92.5% for the South Site) in spite of the difference in the number of sampling units. Thus, the raw data imply that the South Site is more diverse than the North site for a standardized fraction of approximately 92% of the individuals in each assemblage.

In each site, some species were each observed in only one point-count. The existence of such "uniques" signifies that some species were undetected in each site. In the North Site, 9 species were observed in only one point-count ($Q_1 = 9$) and 4 species were observed in two point-counts ($Q_2 = 4$). These 13 rare species contain most of the available information about the number of undetected species. The Chao2 formula in Eq. (2b) implies a species richness estimate for the North Site of 36.3, with a 95% confidence interval of (29.1, 68.8). In the South Site, 12 species were observed in only one point ($Q_1 = 12$), and 8 species were observed in two points ($Q_2 = 8$). The Chao2 formula in Eq. (2b) yields a species richness estimate of 46.5 for the South Site, with a 95% confidence interval of (40.3, 69.8). Richness estimates based on the improved iChao2 estimator (38.6 for North and 48.2 for South), derived by Chiu et al. (2014) in Eq. (6d), differ little from the corresponding Chao2 estimates, so our interpretation is mainly based on the Chao2 estimates. All estimates were computed from the SpadeR Online (Species-richness Prediction And Diversity Estimation Online) software, which is available from Anne Chao's website at `http://chao.stat.nthu.edu.tw/wordpress/software_download/`.

Table 4: *A summary of raw data and species richness estimation for bird species in two sites (the South Site and the North Site in Australian Barrington Tops National Park); see Chao et al. (2015).*

(a) Species incidence frequency counts in the North Site ($S_{obs} = 27$, $T = 12$, total number of incidences $U = 102$, sample coverage estimate $= 91.8\%$); Q_i: the number of species detected in exactly i sampling units (point counts).

i	1	2	3	5	6	7	9	10	11	12
Q_i	9	4	3	5	1	2	1	1	0	1

(b) Species incidence frequency counts in the South Site ($S_{obs} = 38$, $T = 17$, total number of incidences $U = 148$, sample coverage estimate $= 92.5\%$).

i	1	2	3	4	5	6	7	8	9	11	17
Q_i	12	8	3	3	2	2	3	1	1	2	1

(c) Undetected species richness and Chao2 point and interval estimates for each site; see Eq. (2b)

Site	T	Q_1	Q_2	Observed richness	Undetected richness	Chao2 richness	s.e. of Chao2	95% conf. interval
North	12	9	4	27	9.28	36.28	8.31	(29.06, 68.77)
South	17	12	8	38	8.47	46.47	6.43	(40.25, 69.78)

(d) Undetected shared species richness between the two sites and the corresponding Chao2-shared point and interval estimates for shared species richness; see Eq. (10e)

Observed shared richness	Q_{+1}	Q_{+2}	Q_{1+}	Q_{2+}	Q_{11}	Q_{22}
24	6	1	6	4	4	0

$\hat{Q}_{+0}$	$\hat{Q}_{0+}$	$\hat{Q}_{00}$	Undetected shared richness	Chao2-shared richness	s.e. of Chao2-shared	95% conf. interval
2.75	7.06	2.59	12.39	36.39	11.42	(26.67,81.64)

The above results reveal that a relatively high fraction of the species present in each site remain undetected. As discussed in Section 2.3, if we can assume for each site that all undetected and unique species have approximately the same probability to be detected in each point-count, then these asymptotic estimates represent nearly unbiased estimates and can be compared between the two sites. In this case, the data are not sufficient to detect statistically significant differences in richness between the two sites, as reflected by the overlapping confidence intervals associated with the two Chao2 estimates (Table 4). However, the data do support inference of a significance difference in species richness if only a fraction of the assemblages are compared, as shown by the disjoint confidence intervals in the coverage-based rarefaction and extrapolation in the next sub-section.

Table 4 also shows overlap information and shared species richness estimation between the two sites. Out of the 24 observed shared species, 4 were uniques in both sites ($Q_{11} = 4$), 12 shared species were uniques in one site or the other ($Q_{+1} = 6$, $Q_{1+} = 6$), one shared species was a duplicate in the South Site ($Q_{+2} = 1$), and 4 shared species were duplicates in the North Site ($Q_{2+} = 4$). The existence of such rare shared species signifies that there were undetected shared species. Based on the Chao2-shared formula (Eq. 10e), the minimum number of undetected shared species is estimated to be 12.4, and the minimum shared species richness is estimated to be 36.4, with a 95% confidence interval of (26.7, 81.6); see Table 4 for details. Our approach reveals the extent of underestimation and provides helpful information for understanding community/assemblage overlap.

7.3. Rarefaction and extrapolation of species richness (Figures 3, 4 and 5)

We use the data from these two sites to illustrate the construction of two types of rarefaction and extrapolation curves of species richness (sample-size-based and coverage-based), and the sample completeness curve; all formulas are given in Table 2. The constructed sampling curves are then used to compare species richness between the two sites. These sampling curves can be obtained using the online software iNEXT (iNterpolation and EXTrapolation, available from the website address is given in Section 7.2). iNEXT online returns the three sampling curves as shown in Figures 3, 4 and 5, along with some related statistics (omitted here). The omitted output includes basic data information and species richness estimates for some rarefied and extrapolated samples.

The sample-size-based sampling curve (Figure 3) includes a rarefaction part (which plots $\hat{S}(t)$ as a function of $t < T$), and an extrapolation part (which plots $\hat{S}(T+t^*)$ as a function of $T+t^*$), joining smoothly at the reference point $(T,\ S_{obs})$. The confidence intervals based on the bootstrap method also join smoothly. With this type of sampling curve, we can compare species richness for two equally-large samples along with 95% confidence intervals. For each site, the extrapolation is extended to 34 sampling units, double that of the reference sample size of the South Site. Extrapolation beyond the doubled reference sample size could theoretically be computed and used for ranking species richness, but the estimates may be subject to some prediction biases and should be used with caution in estimating species richness ratios or other measures. Figure 3 reveals that the curve for the South Site lies above that of the North Site. However, the confidence intervals of the two sites overlap, implying that comparing two equally-large samples

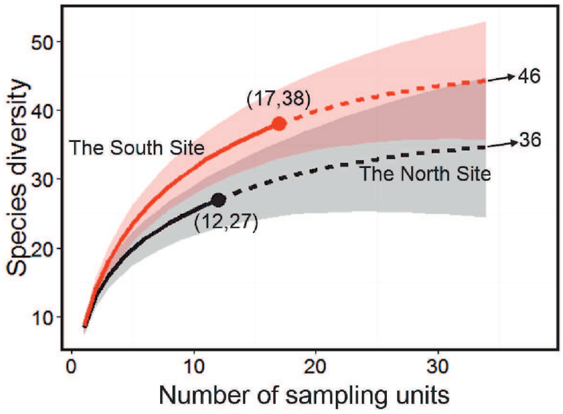


Figure 3: *Sample-size-based rarefaction (solid lines) and extrapolation (dashed lines) sampling curves with 95% confidence intervals (shaded areas, based on a bootstrap method with 200 replications) comparing species richness for Australian bird data in two sites (the South Site and the North Site in Barrington Tops National Park); see Chao et al. (2015). Observed (reference) samples are denoted by the solid dots. The extrapolation extends up to a maximum sample size of 34; here the sample size means the number of sampling units. The numbers in parentheses are the number of sampling units and the observed species richness for each reference sample. The estimated asymptote for each curve is shown next to the arrow at the right-hand end of each curve.*

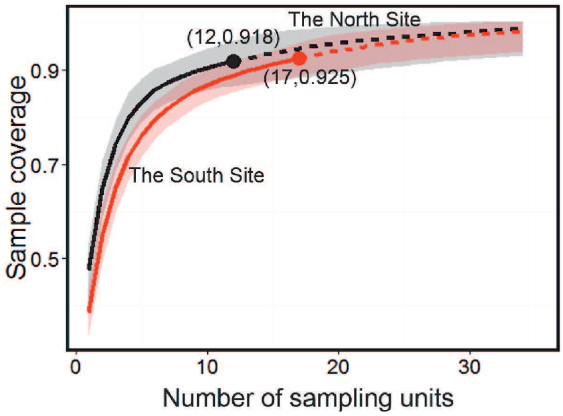


Figure 4: *Plot of sample coverage for rarefied samples (solid line) and extrapolated samples (dashed line) as a function of sample size for Australian bird data in two sites (the South Site and the North Site in Barrington Tops National Park); see Chao et al. (2015). Observed (reference) samples are denoted by solid dots. The 95% confidence intervals (shaded areas) are obtained by a bootstrap method based on 200 replications. Each of the two curves was extrapolated up to the base sample size of 34. The numbers in parentheses are the number of sampling units and the estimated sample coverage for each reference sample.*

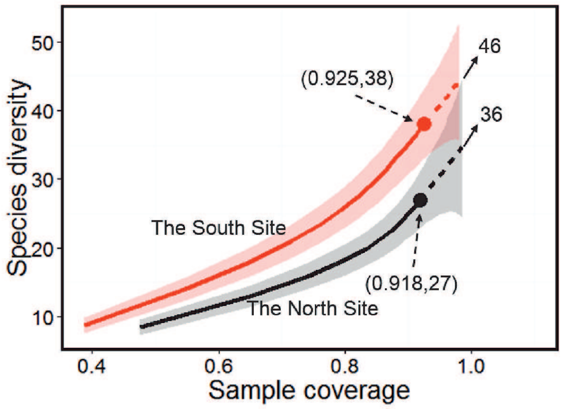


Figure 5: *Coverage-based rarefaction (solid lines) and extrapolation (dashed lines) sampling curves with 95% confidence intervals (shaded areas, based on a bootstrap method with 200 replications) for comparing species richness for Australian bird data in two sites (the South Site and the North Site in Barrington Tops National Park); see Chao et al. (2015). Observed (reference) samples are denoted by solid dots. The extrapolation extends up to the coverage value of the corresponding maximum number of sampling units of 34 in Figure 4 (98.5% in the North Site and 98.1% in the South Site). The numbers in parentheses are the estimated coverage and the observed species richness for each reference sample. The estimated asymptote for each curve is shown next to the arrow at the right-hand end of each curve.*

is inconclusive regarding the test of significant difference in species richness between the two sites. Generally, for any fixed sample size (or completeness) in the comparison range, if the 95% confidence intervals do not overlap, then significant differences at a level of 5% among the expected diversities (whether interpolated or extrapolated) are guaranteed. However, overlapping intervals do not guarantee non-significance (Colwell et al., 2012).

The sample completeness curve (Figure 4) shows how the sample coverage varies with the number of sampling units, along with 95% confidence intervals for each of the two sites, up to the sample size of 34. This curve includes a rarefaction part (which plots $\hat{C}(t)$ as a function of $t < T$), and an extrapolation part (which plots $\hat{C}(T+t^*)$ as a function of $T+t^*$), joining smoothly at the reference point $(T,\hat{C}(T))$. For any fixed number of sampling units, the curve of the North Site lies consistently above that of the South Site, but there is little difference between the two curves when the number of units exceeds 10. For the North Site, when the number of units is extended from 12 to 34, the sample coverage is extended from 91.8% to 98.5% (a number provided by the unreported iNEXT output). For the South Site, when the sample size is extended from 17 to 34 the coverage is extended from 92.5% to 98.1% (as shown in the unreported iNEXT output). The sample completeness curve provides a bridge between sample-size-based and coverage-based sampling curves.

The coverage-based sampling curve (Figure 5) includes a rarefaction part (which plots $\hat{S}(t)$ as a function of $\hat{C}(t)$ for $t < T$), and an extrapolation part (which plots

$\hat{S}(T+t^*)$ as a function of $\hat{C}(T+t^*)$), joining smoothly at the reference sample point ($\hat{C}(T)$, S_{obs}). In this type of sampling curve, we compare species richness for two equally-complete samples along with 95% confidence intervals. The extrapolation is extended to 98.5% for the North Site and to 98.1% for the South Site, as explained in the preceding paragraph. One advantage of using coverage-based curves is that the South Site has significantly greater species richness than the North Site, as evidenced by the non-overlapping confidence intervals for any fixed coverage up to about 93% in Figure 5. This implies that, if we compare species richness for sample coverage up to 93%, the data do provide sufficient information to conclude that the South Site is significantly richer in species. Unlike the sample-sized-based standardization, in which sample size is determined by investigators, the coverage-based standardization compares equal population fractions of each assemblage. The population fraction is an assemblage-level characteristic that can be reliably estimated from data.

As demonstrated in the above-described example, the two R packages (SpadeR and iNEXT) supply useful information for both asymptotic and non-asymptotic analyses. These methods efficiently use all available data to make robust and meaningful comparisons of species richness between assemblages for a wide range of sample sizes/completeness.

7.4. Faith's PD estimation (Table 5)

Without loss of generality, we select the time depth at 82.9 Myr (the age of the root of the phylogenetic tree connecting the observed 41 species) as our temporal perspective for our phylogenetic diversity estimation in this sub-section and for rarefaction/extrapolation in the next sub-section. Although the root of the observed species varies with sampling data, we can easily transform all our estimates to those for a new reference point that is ancestral to all species; see Chao et al. (2015) for transformations.

In the North Site (27 species in 12 sampling units), the observed *PD* (PD_{obs}) is 1222.10 Myr. The total branch lengths for uniques in the sample branch/node incidence frequencies is calculated as $R_1 = 376.5$ Myr, and for duplicates is $R_2 = 153.5$ Myr. These two statistics and the two counts ($Q_1^* = 9$, $Q_2^* = 6$) in the observed tree produce (by Eq. 11c) an estimate of the undetected *PD* as $\hat{R}_0 = 423.3$ Myr, leading to a Chao2-*PD* estimate of the true *PD* of $\widehat{PD}_{Chao2} = PD_{obs} + \hat{R}_0 = 1645.4$, with an estimated s.e. of 465.81 and 95% confidence interval of (1296.0, 3647.9), based on a bootstrap method using 200 replications and a log-transformation.

In the South Site (38 species in 17 sampling units), the observed *PD* (PD_{obs}) is 1416.0 Myr. The corresponding statistics are $R_1 = 376.8$ Myr, $R_2 = 229.5$ Myr, $Q_1^* = 13$ and $Q_2^* = 10$. These yield an estimate of the undetected *PD* as $\hat{R}_0 = 291.2$ Myr, leading to a Chao2-*PD* estimate of the true *PD* of $\widehat{PD}_{Chao2} = PD_{obs} + \hat{R}_0 = 1707.2$, with an estimated s.e. of 206.45 and 95% confidence interval of (1499.4, 2433.1). Thus, signifi-

Table 5: *A summary of phylogenetic data and PD estimation based on the incidence frequency counts data (in Table 4) and the phylogenetic tree (in Figure 2) for bird species in two sites (the South Site and the North Site in Australian Barrington Tops National Park); see Chao et al. (2015). All calculations are based on a reference time point of 82.9 Myr, the age of the root of the phylogenetic tree connecting the observed 41 species.*

Site	Q_1^*	Q_2^*	R_1	R_2	Observed *PD*	Undetected *PD*	Chao2-*PD*	s.e. of Chao2-*PD*	95% conf. interval
North	9	6	376.5	153.5	1222.1	423.3	1645.4	465.81	(1296.0, 3647.9)
South	13	10	376.8	229.5	1416.0	291.2	1707.2	206.45	(1499.4, 2433.1)

cant difference in *PD* between the two sites cannot be guaranteed due to the overlapping confidence intervals.

7.5. Rarefaction and extrapolation of PD (Figures 6 and 7)

The two types of rarefaction and extrapolation curves, along with the sample completeness curves, can be obtained using the online software PhD (Phylogenetic Diversity), available from the website given in Section 7.2. The sample-size-based and coverage-based sampling curves are shown respectively in Figures 6 and 7. These two curves are plotted in the same manner as those for species richness in Section 7.3; the only difference lies in that species richness estimates were replaced by *PD* estimates (all *PD*

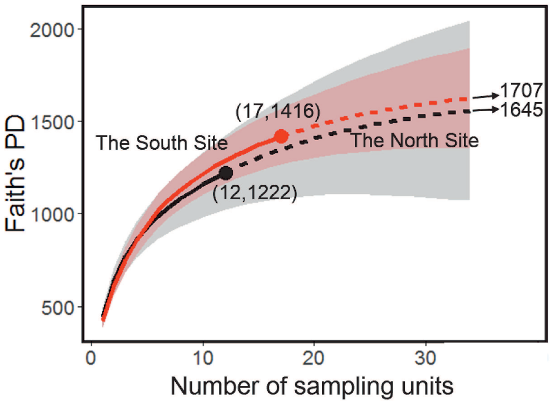


Figure 6: *Comparison of sample-size-based rarefaction (solid lines) and extrapolation (dotted curves) of Faith's PD, up to 34 sampling units for Australian bird data in two sites (the South Site and the North Site in Barrington Tops National Park); see Chao et al. (2015). The fixed time depth is 82.9 Myr (the age of the root of the observed tree.) Observed (reference) samples are denoted by solid dots. The 95% confidence intervals (shaded areas) are obtained by a bootstrap method based on 200 replications. The numbers in parentheses are the number of sampling units and the observed PD for each reference sample. The estimated asymptote of PD (Eq. 11c) for each curve is shown after an arrow sign.*

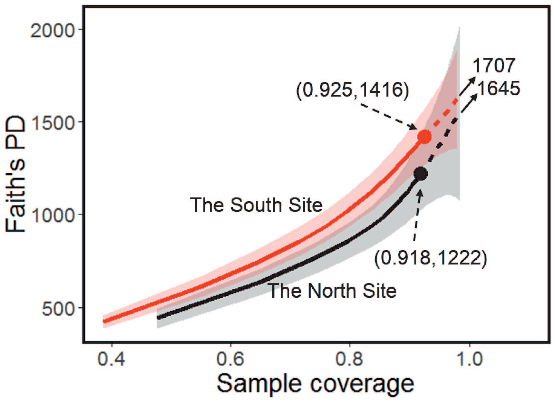


Figure 7: *(a) Comparison of the coverage-based rarefaction (solid lines) and extrapolation (dotted curves) of Faith's PD, up to the coverage 98.5% for the North Site and 98.1% for the South Site for Australian bird data in Barrington Tops National Park (Chao et al., 2015). The fixed time depth is 82.9 Myr (the age of the root of the observed tree.) Observed (reference) samples are denoted by solid dots. The 95% confidence intervals (shaded areas) are obtained by a bootstrap method based on 200 replications. The numbers in parentheses are the estimated sample coverage and the observed PD for each reference sample. The estimated asymptote of PD (Eq. 11c) for each curve is shown after an arrow sign.*

formulas for rarefied and extrapolated samples are provided in the second column of Table 2). The sample completeness curve is identical to that in Figure 4.

We first compare the integrated sample-size-based rarefaction and extrapolation curves for *PD* along with 95% confidence intervals (based on a bootstrap method of 200 replications) up to 34 sampling units. The estimated *PD* and confidence intervals then can be compared across sites for any sample size less than the size of 34. Across this range of sample size, Figure 6 reveals that the South Site has higher *PD* estimate than that of the North Site, but the two confidence intervals overlap and thus data do not provide evidence to support significant difference.

In Figure 7, we compare the corresponding coverage-based rarefaction and extrapolation curves for *PD* with 95% confidence intervals up to the coverage of 98.5% (for the North Site) and 98.1% (for the South Site). Although the estimated *PD* for the South Site still consistently lies above that for the North Site for any standardized sample coverage, the two confidence intervals overlap and thus significant difference cannot be concluded. Chao et al. (2015) analyzed the same data set but based on species abundance data. Although the two types of data yield generally consistent patterns for rarefaction and extrapolation curves, they found that species abundance data show that the PD in the South Site is significantly higher than that in the North-site for any standardized sample coverage less than 90%; see Chao et al. (2015) for analyses based on abundance data.

8. Conclusion and discussion

We have reviewed Chao's (1987) inequality and the associated Chao2 estimator (Eq. 2b) of species richness for multiple incidence data. Using an incidence-data-based generalization of the Good-Turing frequency formula, we have demonstrated that the Chao2 estimator is an unbiased point estimator as long as very rare/infrequent species (specifically, undetected species and unique species in the data) have approximately the same detection probabilities in any sampling unit; the other species (those detected in two or more sampling units) can be highly heterogeneous without affecting the estimator. On the other hand, if very rare/infrequent species are heterogeneous and the sample size is not sufficiently large, then the data do not contain sufficient information to accurately estimate species richness, and the Chao2 formula provides a universal nonparametric lower bound. We have also reviewed the work of Chao et al. (2009) on a related sampling issue, i.e. how many additional sampling units are needed to detect any arbitrary proportion (including 100%) of the Chao2 estimate. Higher-order incidence frequency counts can be also used to construct a series of Chao2-type lower bounds, as derived by Chiu et al. (2014) in Eq. (6d), and by Puig and Kokonendji (2017) in Eq. (6e).

We have also reviewed subsequent developments, including species richness estimators under sampling without replacement, specifically the Chao2-type species richness estimator under sampling without replacement is shown in Eq. (9c). When there are multiple assemblages, the Chao2-shared estimator (Eq. 10d) can be used to infer shared species richness. We also described the Chao2-*PD* estimator (Eq. 11c), which estimates the true *PD* for the phylogenetic tree spanned by all species in the focal assemblage. Similarly, for phylogenetic diversity, the Chao2-*PD* estimator is nearly unbiased if the detection probabilities of rare/infrequent nodes/branches are approximately homogeneous, even if other nodes/branches are heterogeneously detectable. These estimates can be computed from online software SpadeR, iNEXT, and PhD. We have illustrated the use of the software for a real data set in Section 7.

When rare/infrequent species or nodes are highly heterogeneous in their detection probabilities, such as in microbial assemblages or DNA sequencing data, all estimators derived in this paper underestimate the true diversities and can be regarded only as lower bounds. In such cases, a non-asymptotic approach via sample-size-based and coverage-based rarefaction and extrapolation on the basis of standardized sample size or sample completeness (as measured by sample coverage) is recommended. This non-asymptotic approach facilitates fair comparison of diversities (Sections 3 and Section 6.3) for equally-large or equally-complete samples across multiple assemblages. See the example data analysis for rarefaction/extrapolation curves (Figures 3–5 for species richness, and Figures 6 and 7 for *PD*).

None of the diversity measures discussed in this paper (species richness, shared species richness, and Faith's *PD*) considers species abundances. Hill (1973) integrated species richness and species relative abundances into a class of diversity measures later called *Hill numbers*, which include species richness for the diversity order zero. Hill numbers (or the effective number of species) have been increasingly used to quantify the species/taxonomic diversity of assemblages because they represent an intuitive and statistically rigorous alternative to other diversity indices. Hill numbers are parameterized by a diversity order q, which determines the measures' sensitivity to species relative abundances. Hill numbers include the three most widely used species diversity measures as special cases: species richness ($q = 0$), Shannon diversity ($q = 1$), and Simpson diversity ($q = 2$). Like species richness, a Hill number of any order q is dependent on sample size and sample completeness, and thus standardization is needed. The sample-size-based and coverage-based integration of rarefaction (interpolation) and extrapolation (prediction) of Hill numbers represent a unified standardization method for quantifying and comparing species diversity across multiple assemblages; see Chao et al. (2014b) for rarefaction and extrapolation methods based on Hill numbers.

Chao et al. (2010) extended Hill numbers to a class of phylogenetic diversity measures. This class of phylogenetic measures can be regarded as a generalization of Faith's *PD* to incorporate species abundances, because it includes Faith's *PD* as the diversity of order zero ($q = 0$). The corresponding sample-size-based and coverage-based integration of rarefaction and extrapolation of this class of phylogenetic diversity measures was recently developed by Hsieh and Chao (2017). In addition to abundances and evolutionary history, species are often described by a set of traits that affect organismal and/or ecosystem functioning. *Functional diversity* quantifies the diversity of species' traits among coexisting species in an assemblage and is regarded as key to understanding ecosystem processes and their response to environmental stress or disturbance (Tilman et al., 1997; Cadotte et al., 2009). The extension of rarefaction and extrapolation to functional diversity is still under development.

Acknowledgements

We thank the Editor (Pere Puig) for inviting us to write this review paper. We also thank the two reviewers for carefully reading an earlier version and providing very thoughtful and helpful comments and suggestions, which greatly improved the paper. AC was supported by the Taiwan Ministry of Science and Technology under Contracts 105-2628-M007-001.

Appendix A: *Species detection/non-detection records in 12 point-counts for the North Site at Barrington Tops National Park, Australia (Chao et al., 2015).*

Species name	Detection/non-detection record in 12 sampling units (point-counts)												Incidence frequency
Acanthiza_lineata	0	0	0	0	0	0	0	0	0	0	0	0	0
Acanthiza_nana	0	0	0	0	0	0	0	0	0	0	0	0	0
Acanthiza_pusilla	1	1	1	1	1	1	1	1	1	1	1	1	12
Acanthorhynchus_tenuirostris	0	0	0	0	0	0	0	0	0	0	0	0	0
Alisterus_scapularis	0	0	1	0	0	0	0	0	1	0	0	0	2
Cacatua_galerita	0	0	0	0	0	0	1	0	0	0	0	0	1
Cacomantis_flabelliformis	0	1	0	0	1	0	1	1	1	0	0	0	5
Calyptorhynchus_funereus	0	0	0	0	0	0	0	0	0	1	0	0	1
Colluricincla_harmonica	1	1	0	0	0	0	0	0	0	0	0	0	2
Cormobates_leucophaea	1	1	1	1	1	1	1	1	0	1	0	0	9
Corvus_coronoides	0	0	0	0	0	0	0	1	0	0	0	0	1
Dacelo_novaeguineae	0	0	0	0	0	0	0	0	0	0	1	0	1
Eopsaltria_australis	1	0	0	1	0	1	1	0	1	0	0	0	5
Gerygone_mouki	1	1	1	1	0	1	0	0	0	0	0	0	5
Leucosarcia_melanoleuca	0	1	0	0	0	0	0	0	0	0	0	0	1
Lichenostomus_chrysops	0	0	0	0	0	0	0	0	0	0	0	0	0
Malurus_cyaneus	0	0	0	0	0	0	0	0	0	0	0	0	0
Malurus_lamberti	0	0	0	0	0	0	0	0	0	0	0	0	0
Manorina_melanophrys	0	0	0	0	0	0	0	0	0	0	0	0	0
Meliphaga_lewinii	1	1	1	1	1	1	0	0	0	0	0	0	6
Menura_novaehollandiae	1	1	1	0	0	0	0	0	1	0	0	1	5
Monarcha_melanopsis	0	0	0	0	1	0	0	0	0	0	0	0	1
Neochmia_temporalis	0	0	0	0	0	0	0	0	0	0	0	0	0
Oriolus_sagittatus	0	0	1	0	0	0	0	0	0	0	0	0	1
Pachycephala_olivacea	0	0	0	0	0	0	0	0	0	0	0	0	0
Pachycephala_pectoralis	1	1	1	1	0	1	0	0	0	0	1	1	7
Pachycephala_rufiventris	0	0	0	0	0	0	0	0	0	0	0	0	0
Pardalotus_punctatus	1	0	1	1	1	0	1	1	0	0	1	0	7
Petroica_rosea	0	0	0	0	0	0	0	0	0	0	1	0	1
Phylidonyris_niger	0	0	0	0	0	0	0	0	0	0	0	0	0
Platycercus_elegans	0	0	1	0	0	0	0	0	0	0	0	0	1
Psophodes_olivaceus	0	0	1	1	0	0	0	1	0	0	0	0	3
Ptilonorhynchus_violaceus	0	0	0	0	1	0	0	0	0	0	0	1	2
Ptiloris_paradiseus	0	0	0	0	0	0	0	0	0	0	0	0	0
Rhipidura_albicollis	1	1	0	1	0	1	1	1	1	1	1	1	10
Rhipidura_rufifrons	0	0	1	0	1	0	0	0	0	1	0	0	3
Sericornis_citreogularis	0	0	0	0	0	0	0	0	0	0	0	0	0
Sericornis_frontalis	1	0	0	1	0	0	0	0	0	0	0	0	2
Strepera_graculina	0	0	1	0	0	0	0	0	0	1	0	1	3
Zoothera_lunulata	0	0	0	0	0	0	0	0	0	0	0	0	0
Zosterops_lateralis	1	1	0	0	1	1	1	0	0	0	0	0	5

Appendix B: *Species detection/non-detection records in 17 point-counts for the South Site at Barrington Tops National Park, Australia (Chao et al., 2015).*

Species name	Detection/non-detection record in 17 sampling units (point-counts)																	Incidence frequency
Acanthiza_lineata	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Acanthiza_nana	0	0	0	1	0	1	1	0	0	1	0	0	0	0	1	0	0	5
Acanthiza_pusilla	0	0	0	0	1	0	0	1	0	0	1	1	1	1	1	1	1	9
Acanthorhynchus_tenuirostris	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	2
Alisterus_scapularis	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Cacatua_galerita	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Cacomantis_flabelliformis	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	4
Calyptorhynchus_funereus	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Colluricincla_harmonica	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	4
Cormobates_leucophaea	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17
Corvus_coronoides	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dacelo_novaeguineae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Eopsaltria_australis	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	3
Gerygone_mouki	0	0	0	0	1	1	1	1	0	0	1	0	1	1	0	0	0	7
Leucosarcia_melanoleuca	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Lichenostomus_chrysops	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Malurus_cyaneus	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2
Malurus_lamberti	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	2
Manorina_melanophrys	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Meliphaga_lewinii	0	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	0	11
Menura_novaehollandiae	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	4
Monarcha_melanopsis	0	0	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	5
Neochmia_temporalis	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	2
Oriolus_sagittatus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pachycephala_olivacea	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	2
Pachycephala_pectoralis	0	0	0	0	1	0	1	0	0	1	0	0	1	1	1	1	0	7
Pachycephala_rufiventris	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Pardalotus_punctatus	0	0	1	1	0	1	1	1	1	0	0	0	0	0	1	0	1	8
Petroica_rosea	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
Phylidonyris_niger	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Platycercus_elegans	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2
Psophodes_olivaceus	0	1	0	0	0	0	1	1	0	0	1	1	0	0	0	1	0	6
Ptilonorhynchus_violaceus	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
Ptiloris_paradiseus	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	2
Rhipidura_albicollis	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	1	1	11
Rhipidura_rufifrons	0	0	0	1	0	1	1	1	0	1	1	1	0	0	0	0	0	7
Sericornis_citreogularis	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
Sericornis_frontalis	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	3
Strepera_graculina	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3
Zoothera_lunulata	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Zosterops_lateralis	0	1	0	0	0	1	0	1	0	1	0	0	0	1	0	0	1	6

References

Böhning, D. and van der Heijden, P.G.M. (2009). A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *Annals of Applied Statistics*, 3, 595–610.

Böhning, D., Vidal-Diez, A., Lerdsuwansri, R., Viwatwongkasem, C. and Arnold, M. (2013). A Generalization of Chao's estimator for covariate information. *Biometrics*, 69, 1033–1042.

Cadotte, M.W., Cavender-Bares, J., Tilman, D. and Oakley, T.H. (2009). Using phylogenetic, functional and trait diversity to understand patterns of plant community productivity. *PLoS One*, 4, e5695.

Cardoso, P., Rigal, F., Borges, P.A. and Carvalho J.C. (2014). A new frontier in biodiversity inventory: a proposal for estimators of phylogenetic and functional diversity. *Methods in Ecology and Evolution*, 5, 452–461.

Cavender-Bares, J., Ackerly, D.D. and Kozak, K.H. (2012). Integrating ecology and phylogenetics: the footprint of history in modern-day communities. *Ecology*, 93, S1–S3.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11, 265–270.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43, 783–791.

Chao, A. and Lee, S-M. (1992). Estimating the number of classes via sample coverage. *Journal of American Statistical Association*, 87, 210–217.

Chao, A., Lee, S-M. and Jeng, C-L. (1992). Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, 48, 201–216.

Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological and Environmental Statistics*, 6, 158–175.

Chao, A. (2005). Species estimation and applications. In Balakrishnan, N., C. Read, B, and B. Vidakovic (eds) *Encyclopedia of Statistical Sciences*, 7907–7916. Wiley, New York.

Chao, A., Chazdon, R.L., Colwell, R.K. and Shen, T.-J. (2005). A new statistical approach for assessing compositional similarity based on incidence and abundance data. *Ecology Letters*, 8, 148–159.

Chao, A., Chazdon, R.L., Colwell, R.K. and Shen, T.-J. (2006). Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics*, 62, 361–371.

Chao, A., Colwell, R.K., Lin, C.-W. and Gotelli, N.J. (2009). Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*, 90, 1125–1133.

Chao, A., Chiu, C.-H. and Jost, L. (2010). Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365, 3599–3609.

Chao, A. and Chiu, C.-H. (2012). Estimation of species richness and shared species richness. In: Balakrishnan, N (ed) *Methods and Applications of Statistics in the Atmospheric and Earth Sciences*, 76–111. Wiley, New York.

Chao, A. and Jost, L. (2012). Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93, 2533–2547.

Chao, A. and Lin, C.-W. (2012). A nonparametric lower bound for species richness and shared species richness under sampling without replacement. *Biometrics*, 68, 912–921.

Chao, A., Chiu, C.-H. and Jost, L. (2014a). Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. *Annual Reviews of Ecology, Evolution, and Systematics*, 45, 297–324.

Chao, A., Gotelli, N.J., Hsieh, T., Sander, E.L., Ma, K., Colwell, R.K. and Ellison, A.M. (2014b). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, 84, 45–67.

Chao, A., Chiu, C.-H., Hsieh, T., Davis, T., Nipperess, D.A. and Faith, D.P. (2015). Rarefaction and extrapolation of phylogenetic diversity. *Methods in Ecology and Evolution*, 6, 380–388.

Chao, A. and Chiu, C.-H. (2016). Species richness: estimation and comparison. *Wiley StatsRef: Statistics Reference Online*. 1–26.

Chao, A., Chiu, C.-H., Colwell, R.K., Magnago, L.F.S., Chazdon, R.L. and Gotelli, N.J. (2017). Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on Good-Turing theory. *Ecology*, under revision.

Chazdon, R.L., Colwell, R.K., Denslow, J.S. and Guariguata, M.R. (1998). Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of Northeastern Costa Rica. In: Dallmeier, F., Comiskey, J.A. (eds.). *Forest Biodiversity Research, Monitoring and Modeling: Conceptual Background and Old World Case Studies*. 285–309. Parthenon Publishing, Paris.

Chiarucci, A., Bacaro, G., Rocchini, D. and Fattorini, L. (2008). Discovering and rediscovering the sample-based rarefaction formula in the ecological literature. *Community Ecology*, 9, 121–123.

Chiu, C.H., Wang, Y.T., Walther, B.A. and Chao, A. (2014). An improved nonparametric lower bound of species richness via a modified Good–Turing frequency formula. *Biometrics*, 70, 671–682.

Coleman, B.D., Mares, M.A., Willig, M.R. and Hsieh, Y.H. (1982). Randomness, area, and species richness. *Ecology*, 63, 1121–1133.

Colwell, R.K. and Coddington, J.A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B - Biological Sciences*, 345, 101–118.

Colwell, R.K., Mao, C.X. and Chang, J. (2004). Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, 85, 2717–2727.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.-Y., Mao, C.X., Chazdon, R.L. and Longino, J.T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5, 3–21.

Colwell, R.K. (2013). EstimateS: Statistical estimation of species richness and shared species from samples. Version 9 and earlier. User's Guide and application. Published at: `http://purl.oclc.org/estimates`.

Colwell, R.K. and Elsensohn, J.E. (2014). EstimateS turns 20: statistical estimation of species richness and shared species from samples, with non-parametric extrapolation. *Ecography*, 37, 609–613.

Crozier, R.H. (1997). Preserving the information content of species: genetic diversity, phylogeny, and conservation worth. *Annual Review of Ecology and Systematics*, 28, 243–268.

Faith, D.P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61, 1–10.

Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40, 237–264.

Good, I.J. (1983). *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesota Press.

Good, I.J. (2000). Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval enigma. *Journal of Statistical Computation and Simulation*, 66, 101–111.

Good, I.J. and Toulmin G. (1956). The number of new species and the increase of population coverage when a sample is increased. *Biometrika*, 43, 45–63.

Gotelli, N.J. and Chao, A. (2013). Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In Levin, S.A. (Ed). *Encyclopedia of Biodiversity*, 2$^{\text{nd}}$ Edition, Vol. 5, 195–211, Waltham, MA: Academic Press.

Gotelli, N.J. and Colwell, R.K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, 4, 379–391.

Gotelli, N.J. and Colwell, R.K. (2011). Estimating species richness. In: A. Magurran and B. McGill (eds). *Biological Diversity: Frontiers in Measurement and Assessment*, 39–54. Oxford University Press, Oxford.

Heck, K.L., Jr., van Belle G. and Simberloff, D. (1975). Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology*, 56, 1459–61.

Hill, M. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54, 427–432.

Hsieh, T.C. and Chao, A. (2017). Rarefaction and extrapolation: making fair comparison of abundance-sensitive phylogenetic diversity among multiple assemblages. *Systematic Biology*, 66, 100–111.

Hughes, G. and Madden, L.P. (1993). Using the beta-binomial distribution to discrete aggregated patterns of disease incidence. *Phytopathology*, 83, 759–763.

Hurlbert, S.H. (1971). The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, 52, 577–586.

Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K. and Mooers, A.O. (2012). The global diversity of birds in space and time. *Nature*, 491, 444–448.

Jost, L., Chao, A. and Chazdon, R. (2011). Compositional similarity and beta diversity. In A. Magurran and B. McGill (eds). *Biological Diversity: Frontiers in Measurement and Assessment*, 66–84. Oxford University Press, Oxford.

Lanumteang, K. and Böhning, D. (2011). An extension of Chao's estimator of population size based on the first three capture frequency counts. *Computational Statistics & Data Analysis*, 7, 2302–2311.

Magurran, A.E. (2004). *Measuring Biological Diversity*. Blackwell Publishing, Oxford.

Magurran, A.E. and McGill, B.J. (eds) (2011). *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press, Oxford.

Mao, C.X. (2006). Inference on the number of species through geometric lower bounds. *Journal of American Statistical Association*, 101, 1663–1670.

Mao, C.X. (2008). Lower bounds to the population size when capture probabilities vary over individuals. *Australian and New Zealand Journal of Statistics*, 50, 125–134.

Mao, C.X. and Lindsay, B.G. (2007). Estimating the number of classes. *Annals of Statistics*, 35, 917–930.

Mao, C.X., Yang, N. and Zhang, J. (2013). On population size estimators in the Poisson mixture model. *Biometrics*, 69, 758–765.

McGrayne, S.B. (2011). The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy. *Yale University Press*, New Haven, Connecticut.

Pan, H.Y., Chao, A. and Foissner, W. (2009). A non-parametric lower bound for the number of species shared by multiple communities. *Journal of Agricultural, Biological and Environmental Statistics*, 14, 452–468.

Petchey, O.L. and Gaston, K.J. (2002). Functional diversity (FD), species richness and community composition. *Ecology Letters*, 5,402–411.

Puig, P. and Kokonendji, C. (2017). Nonparametric estimation of the number of zeros in truncated count distributions. To appear in the *Scandinavian Journal of Statistics*.

Rivest, L.P. and Baillargeon, S. (2007). Applications and extensions of Chao's moment estimator for the size of a closed population. *Biometrics*, 63, 999–1006.

Sanders, H.L. (1968). Marine benthic diversity: a comparative study. *American Naturalist*, 102, 243–282.

Shinozaki, K. (1963). Notes on the species-area curve, *10th Annual Meeting of the Ecological Society of Japan* (Abstract), p. 5.

Shiyomi, M., Takahashi, S. and Yoshimura, J. (2000). A measure for spatial heterogeneity of a grassland vegetation based on the beta-binomial distribution. *Journal of Vegetation Science*, 11, 627–632.

Simberloff, D. (1972). Properties of the rarefaction diversity measurement. *American Naturalist*, 106, 414–418.

Skellam, J.G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society B*, 10, 257–261.

Tilman, D., Knops, J., Wedin, D., Reich, P., Ritchie, M. and Siemann, E. (1997). The influence of functional diversity and composition on ecosystem processes. *Science*, 277, 1300–1302.

Warwick, R.M. and Clarke, K.R. (1995). New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. *Marine Ecology Progress Series*, 129, 301–305.

Webb, C.O. and Donoghue, M.J. (2005). Phylomatic: tree assembly for applied phylogenetics. *Molecular Ecology Notes*, 5, 181–183.

SORT 41 (1) January-June 2017, 55-72 DOI: 10.2436/20.8080.02.50

On a property of Lorenz curves with monotone elasticity and its application to the study of inequality by using tax data

M.A. Sordo, A. Berihuete, C.D. Ramos and H.M. Ramos[*]

Abstract

The Lorenz curve is the most widely used graphical tool for describing and comparing inequality of income distributions. In this paper, we show that the elasticity of this curve is an indicator of the effect, in terms of inequality, of a truncation of the income distribution. As an application, we consider tax returns as equivalent to the truncation from below of a hypothetical income distribution. Then, we replace this hypothetical distribution by the income distribution obtained from a general household survey and use the dual Lorenz curve to anticipate this effect.

MSC: 91B02.

Keywords: Lorenz curve, tax data, truncation, inequality.

1. Introduction

Tax data are commonly used sources of information in the analysis of income distributions. For example, Piketty and Saez (2003) used tax data to study the concentration of income within the top 10 percent of the distribution with higher incomes in the United States and regularly release reports with the latest available data. Atkinson, Piketty and Saez (2011) provided a comparative study of top incomes covering a wide variety of countries by using tax data. More recently, Saez and Zucman (2014) expanded these works to examine trends in wealth concentration. Tax returns, like other administrative sources, often provide more accurate and complete data for the population under study than other surveys (see Stone et al., 2015).

Research based on data from income tax returns focuses on people who file taxes. However, not everyone is required to file an income tax return every year. In general,

[*] Dpto. de Estadística e I.O., Universidad de Cádiz, Spain. mangel.sordo@uca.es, angel.berihuete@uca.es, carmen.ramos@uca.es, hector.ramos@uca.es
Received: September 2015
Accepted: November 2016

people who do not file are those whose income falls below certain thresholds. The exclusion of non-filers is the most significant limitation of this source if the target population is the entire population in a country. In this case, data from tax returns produce an obvious systematic error or bias in estimating characteristics related to the size of income, such as the disposable income or per capita income. The question that we investigate in this paper is whether there are a systematic error in evaluating, from tax data, characteristics related to the inequality of the income distribution for the entire population. The approach adopted here is to consider tax returns as equivalent to a truncation from below on a hypothetical distribution that would be obtained if everyone would pay taxes. Since this hypothetical distribution is unrealistic, in practice we replace it by the income distribution obtained from a general household survey. It is shown that the effect of income truncation (at any level) by itself does not necessarily introduce a bias in one direction or the other and that when it does, it depends on the shape of the Lorenz curve associated to this distribution.

Specifically, let X be the random variable that describes the true income distribution, let F be its cumulative distribution function and assume that X has a finite mean $\mu > 0$. The most widely used graphical tool for describing and comparing inequality of income distributions is the Lorenz curve[1]. For the income random variable X, the Lorenz curve is defined by

$$L(p) = \frac{1}{\mu}\int_0^p F^{-1}(t)dt, \qquad 0 \leq p \leq 1 \tag{1}$$

where we denote by F^{-1} the inverse of F defined by $F^{-1}(t) = \inf\{x : F(x) \geq t\}$, with $0 < t < 1$. For each p in $(0,1)$, the function $L(p)$ is the cumulative percetage of total income held by individuals having the $100p\%$ lowest incomes. In this paper, the main result is depicted in terms of the dual Lorenz curve $\overline{L}(p)$, which is a reverse-mirror image of the Lorenz curve reflected through the diagonal 45 degree line. It is defined by

$$\overline{L}(p) = 1 - L(1-p),\ 0 \leq p \leq 1, \tag{2}$$

and represents the proportion of total income that accrues to individuals having the $100p\%$ largest incomes. Both curves are non-decreasing and differentiable almost everywhere, with $L(0) = \overline{L}(0) = 0$ and $L(1) = \overline{L}(1) = 1$. If the distribution function F is continuous and strictly increasing, then $L(p)$ and $\overline{L}(p)$ are strictly increasing and continuously differentiable functions of p. The Lorenz curve induces a partial ordering (denoted $\leq_L$, see Arnold, 1987) on the class of income random variables, ordering them in terms of inequality. Given two income random variables X and Y with Lorenz curves $L_X(p)$ and $L_Y(p)$, respectively, X is less unequal than Y (denoted by $X \leq_L Y$) if and only

[1] Lorenz curves are used in many diverse fields, other than income distributions, such as informetrics (see Sarabia, Prieto and Trueba, 2012), demography (see Ramos et al., 2013) or risk measurement (see Greselin and Zitikis, 2015) among others.

if

$$L_X(p) \geq L_Y(p) \text{ for all } 0 \leq p \leq 1 \tag{3}$$

or, equivalently, if

$$\overline{L}_X(p) \leq \overline{L}_Y(p) \text{ for all } 0 \leq p \leq 1.$$

Under a progressive tax structure there exists a tax-free threshold t below which people do not pay personal income tax. Thus, tax return distribution is obtained by a lower truncation of X at income t. For each $t > 0$, denote by $X_{(t,\infty)} = \{X \mid X \geq t\}$ the corresponding lower truncated random variable. In Section 2, we give conditions to compare, in terms of inequality, the income random variables X and $X_{(t,\infty)}$, without needing to know the distribution function of X. Thus, unlike other authors that have previously studied this topic (Bhattacharya, 1963; Moothathu, 1986; Belzunce, Candel and Ruiz, 1995, 1998), our conditions can be directly verified from the Lorenz curve of X. As we explain in Section 2, our results can also be stated in terms of the monotonicity of the function

$$e(t) = E\left[\frac{X}{t}|X > t\right]$$

which represents the expected proportional income to t, for incomes greater than t.

Inequality is not the only characteristic of interest of income distributions. Another important aspect of the concentration of incomes is related to the notion of relative deprivation, which is based on the perception that an individual makes about his social status in a population. In order to compare distributions in terms of deprivation, the starshaped order (see Shaked and Shantikumar, 2007) and the expected proportional shortfall order (Belzunce et al., 2012, 2013) can be considered. In Section 3 we study the effect of truncations on these orderings. It will be shown that the effect of truncations on the expected proportional shortfall order depends, like in the case of the Lorenz order, on the elasticity of the Lorenz curve. In Section 4 we review some parametric models for the Lorenz curve that satisfy the conditions stated in previous sections. Finally, in Section 5, we illustrate the usefulness of our results by a descriptive study based on real data drawn from the survey EU-SILC 2010.

2. Lorenz ordering of truncated random variables

We show in this section that the effect of truncation on the inequality depends on the sensitivity of the dual Lorenz $\overline{L}(p)$ with respect to a change in p, that is, on its elasticity $\varepsilon_{\overline{L}}(p)$, defined by

$$\varepsilon_{\overline{L}}(p) = \frac{d\log\overline{L}(p)}{d\log p} = \frac{pL'(1-p)}{1-L(1-p)},\ 0 < p < 1. \tag{4}$$

We have the following result.

Theorem 2.1. *Let X be an income random variable with Lorenz curve L and let $t_0 \in [0,1]$. Then*

$$X_{(t,\infty)} \leq_L X_{(t',\infty)} \ (\geq_L) \ \textit{for all } t_0 < t < t' \tag{5}$$

if and only if $\varepsilon_{\overline{L}}(p)$ is increasing (decreasing) in the interval $(0, 1-F(t_0))$.

Proof. We give the proof for the case $\leq_L$ (the proof for the $\geq_L$ case is analogous). For each $t > 0$, denote by $F_{(t,\infty)}(x)$ the distribution function of $X_{(t,\infty)}$ given by

$$F_{(t,\infty)}(x) = \begin{cases} 0 & x < t \\ \frac{F(x)-F(t)}{1-F(t)} & x \geq t \end{cases}. \tag{6}$$

Let $L_{(t,\infty)}(p)$ be the Lorenz curve of $X_{(t,\infty)}$. By using (1) and (6) it is easy to see that

$$L_{(t,\infty)}(p) = \frac{L[(1-F(t))p+F(t)] - L(F(t))}{1-L(F(t))}, \ 0 \leq p \leq 1. \tag{7}$$

Condition (5) holds if and only if $L_{(t,\infty)}(p)$ is decreasing in $t > t_0$ or, equivalently, if

$$1 - L_{(t,\infty)}(p) \text{ is increasing in } t > t_0.$$

By making $F(t) = a$ and using (7), this is the same as

$$\frac{1 - L[(1-a)p+a]}{1-L(a)} \text{ is increasing in } a > F(t_0)$$

which, by differentiation, is satisfied if and only if

$$\frac{(1-p)L'[(1-a)p+a]}{1-L[(1-a)p+a]} \leq \frac{L'(a)}{1-L(a)}, \ p \in [0,1], \ a > F(t_0).$$

The above inequality can be rewritten as

$$\frac{(1-a)(1-p)L'[(1-a)p+a]}{1-L[(1-a)p+a]} \leq \frac{(1-a)L'(a)}{1-L(a)}, \ \ p \in [0,1], \ a > F(t_0),$$

which, by making $p_1 = (1-a)(1-p)$ and $p_2 = 1-a$, is the same as

$$\frac{p_1 L'(1-p_1)}{1-L(1-p_1)} \leq \frac{p_2 L'(1-p_2)}{1-L(1-p_2)} \text{ whenever } 0 < p_1 \leq p_2 < 1 - F(t_0).$$

Using (4), this means that $\varepsilon_{\overline{L}}(p)$ is increasing in the interval $(0, 1-F(t_0))$. ■

By taking $t_0 = 0$ we obtain the following corollary.

Corollary 2.2. *Let X be an income random variable with Lorenz curve L. If $\varepsilon_{\overline{L}}(p)$ is increasing (decreasing), then*

$$X \leq_L X_{(t,\infty)}\ (\geq_L)\ \textit{for all}\ t.$$

Using (1) and (4), we see that the elasticity of the dual Lorenz curve can be written as

$$\varepsilon_{\overline{L}}(p) = \frac{F^{-1}(1-p)}{E\left[X\,|X > F^{-1}(1-p)\right]},\ 0 < p < 1. \tag{8}$$

Thus, the increasing (or decreasing) monotonicity of $\varepsilon_{\overline{L}}(p)$ in the interval $(0, 1-F(t_0))$ is equivalent to the increasing (respectively, decreasing) monotonicity of the function

$$e(t) = E\left[\frac{X}{t}\,|X > t\right]$$

in the interval (t_0, ∞). For an income t, the function $e(t)$ represents the expected proportional income to t, for incomes greater than t. This function was used by Belzunce, Candel and Ruiz (1998) to characterize the effect of truncation of a random variable X on the Lorenz curve. They say that X is DMLPRI (decreasing mean proportional residual income) if $e(t)$ is decreasing in t. From the above observation, we can equivalently say that X is DMLPRI if $\varepsilon_{\overline{L}}(p)$ is decreasing. It is worth noting that Theorem 2.1 and the rest of results in this paper involving $\varepsilon_{\overline{L}}(p)$ can be easily reformulated in terms of the curve $e(t)$.

Bhattacharya (1963) showed that the Lorenz curve of a lower truncated income distribution is independent of the point of truncation if, and only if, the incomes follow the Pareto law, with distribution function

$$F(x) = 1 - \left(\frac{\theta}{x}\right)^{a},\ \theta > 0, a > 0, x > \theta. \tag{9}$$

Now, combining the result of Bhattacharya with Theorem 2.1, we can characterize the Pareto distribution in terms of the elasticity of the dual Lorenz curve.

Corollary 2.3. *Let X be an income random variable with Lorenz curve L. Then, X follows the Pareto distribution if and only if $\overline{L}(p)$ has a constant elasticity.*

Similar results can be stated for upper truncations. If we denote by $X_{(0,s)} = \{X \mid X \leq s\}$ the upper truncated random variable at income s, it can be shown that the corresponding Lorenz curve $L_{(0,s)}(p)$ satisfies

$$L_{(0,s)}(p) = \frac{L(F(s)p)}{L(F(s))},\ 0 \leq p \leq 1. \tag{10}$$

The comparison of upper truncations of a random variable X is characterized in terms of the elasticity of the Lorenz curve

$$\varepsilon_L(p) = \frac{d\log L(p)}{d\log p} = \frac{pL'(p)}{L(p)},\ 0 < p < 1.$$

The proof of the following result follows the same lines as the proof of Theorem 2.1 and therefore it is omitted.

Theorem 2.4. *Let X be an income random variable with Lorenz curve L and let $s_0 \in [0,1]$. Then*

$$X_{(0,s)} \leq_L X_{(0,s')}\ (\geq_L)\, \text{for all } s < s' < s_0 \tag{11}$$

if and only if $\varepsilon_L(p)$ is increasing (decreasing) in $[F(s_0),1]$.

Moothathu (1986) showed that the Lorenz curve is unchanged by upper truncation if, and only if, incomes follow a power law, with distribution function

$$F(x) = \left(\frac{x}{\lambda}\right)^a,\ \lambda > 0, a > 0, 0 < x < \lambda. \tag{12}$$

The combination of this result with Theorem 2.4 let us characterize the power distributions in terms of the elasticity of the Lorenz curve.

Corollary 2.5. *Let X be an income random variable with Lorenz curve L. Then, X follows the power distribution if and only if $L(p)$ has a constant elasticity.*

3. The effect of truncations on the starshaped order and the expected proportional shortfall order

In Section 2 we have shown that the effect of truncation on the inequality depends on the elasticities of the Lorenz curve $L(p)$ and its dual $\overline{L}(p) = 1 - L(p)$. However, the Lorenz curve is not the only tool for comparing income distributions in terms of concentration. The Lorenz order is a pure inequality order, in the sense that it is consistent with the well-known Pigou–Dalton Transfer Principle, which demands that a transfer from a richer person to a poorer person of less than the difference in their income unambiguously reduces inequality. When we compare income distributions in terms of relative status of people or relative deprivation (rather than in terms of inequality), some other orderings, such as the starshaped order and the expected proportional shortfall order, can also be considered (see Shaked and Shantikumar, 2007, and Belzunce et al., 2012, 2013, 2016, for properties and applications of these orders) and it is of interest to investigate whether similar results for truncated distributions can be given taking into account the elasticity of some related functions like, for example, the quantile function.

First, we define these orders.

Definition 3.1. *Given two income random variables X and Y, with distribution functions F and G, respectively, then:*

(i) We say that X is smaller than Y in the starshaped order (denoted by $X \leq_* Y$) if

$$\frac{G^{-1}(p)}{F^{-1}(p)} \text{ is increasing in } p \in (0,1)\,.$$

(ii) We say that X is smaller than Y in the expected proportional shortfall order (denoted by $X \leq_{ps} Y$) if

$$E\left[\left(\frac{X - F^{-1}(p)}{F^{-1}(p)}\right)^+\right] \leq E\left[\left(\frac{Y - G^{-1}(p)}{G^{-1}(p)}\right)^+\right] \text{ for } p \in (0,1)\,,$$

where $(x)^+ = x$ if $x \geq 0$ and $(x)^+ = 0$ if $x < 0$.

It can be shown (see Theorem 2.11 in Belzunce et al., 2012) that $X \leq_{ps} Y$ if and only if

$$\int_p^1 \left[\frac{F^{-1}(t)}{F^{-1}(p)}\right] dt \leq \int_p^1 \left[\frac{G^{-1}(t)}{G^{-1}(p)}\right] dt, \text{ for all } p \in (0,1)\,. \tag{13}$$

On the other hand, it is well-known that

$$X \leq_* Y \Longrightarrow X \leq_{ps} Y \Longrightarrow X \leq_L Y.$$

The next result shows that the effect of truncations on the starshaped order depends on the the elasticities of the quantile function and the inverse of the survival function.

Theorem 3.2. *Let X be an absolutely continuous income random variable with distribution function F and survival function $\overline{F} = 1 - F$. Denote by $\varepsilon_{F^{-1}}(p)$ the elasticity of the quantile function $F^{-1}(p)$ and by $\varepsilon_{\overline{F}^{-1}}(p)$ the elasticity of the inverse survival function $\overline{F}^{-1}(p)$. Then*

(i) $X_{(t,\infty)} \leq_* X_{(t',\infty)}$ $(\geq_*)$ for all $t < t'$ if and only if $\varepsilon_{\overline{F}^{-1}}(p)$ is increasing (decreasing) in $p \in (0,1)$.

(ii) $X_{(0,s)} \leq_* X_{(0,s')}$ $(\geq_*)$ for all $s < s'$ if and only if $\varepsilon_{F^{-1}}(p)$ is increasing (decreasing) in $p \in (0,1)$.

Proof. Let f be density function of X. In order to prove (i), observe that

$$\varepsilon_{\overline{F}^{-1}}(p) = \frac{-p}{f\left(\overline{F}^{-1}(p)\right)\overline{F}^{-1}(p)}, \text{ for } p \in (0,1)\,.$$

Belzunce, Candel and Ruiz (1995, Theorem 3) showed that $X_{(t,\infty)} \leq_{*} X_{(t',\infty)}$ $(\geq_{*})$ for all $t < t'$ if and only if the function

$$\frac{xf(x)}{1-F(x)} \text{ is decreasing (increasing).} \tag{14}$$

By making the change $x = \overline{F}^{-1}(p) = F^{-1}(1-p)$, we see that (14) is equivalent to say that

$$\frac{\overline{F}^{-1}(p) f\left(\overline{F}^{-1}(p)\right)}{p} \text{ is increasing (decreasing),}$$

which holds if, and only if, $\varepsilon_{\overline{F}^{-1}}(p)$ is increasing (decreasing). Part (ii) is proven similarly by using Theorem 4 of Belzunce, Candel and Ruiz (1995). ■

Next we show that the effect of trunctions on the expected proportional shortfall order depends, like in the case of the Lorenz order, on the elasticities of the Lorenz curve an its dual.

Theorem 3.3. *Let X be an absolutely continuous income random variable with distribution function F and survival function $\overline{F} = 1 - F$. Then*

(i) $X_{(t,\infty)} \leq_{ps} X_{(t',\infty)}$ $(\geq_{ps})$ for all $t < t'$ if and only if $\varepsilon_{\overline{L}}(p)$ is increasing (decreasing) in $p \in (0,1)$.

(ii) $X_{(0,s)} \leq_{ps} X_{(0,s')}$ $(\geq_{ps})$ for all $s < s'$ if and only if $\varepsilon_{L}(p)$ is increasing (decreasing) in $p \in (0,1)$.

Proof. We only prove the case $\leq_{ps}$ of part (i), the case $\geq_{ps}$ and part (ii) are proven similarly. First observe from (8) that $\varepsilon_{\overline{L}}(p)$ can be written as

$$\varepsilon_{\overline{L}}(p) = \frac{pF^{-1}(1-p)}{\int_{1-p}^{1} F^{-1}(t)dt}, \quad 0 < p < 1. \tag{15}$$

Let $F_{(t,\infty)}(x)$ be the distribution function of $X_{(t,\infty)}$ given by (6) and let

$$F^{-1}_{(t,\infty)}(u) = F^{-1}\left[(1-F(t))\,u + F(t)\right],\; u \in (0,1), \tag{16}$$

be the corresponding quantile function. Suppose that

$$X_{(t,\infty)} \leq_{ps} X_{(t',\infty)} \text{ for all } t < t'$$

or equivalently, using (13), that

$$\int_p^1 \left[\frac{F^{-1}_{(t,\infty)}(u)}{F^{-1}_{(t,\infty)}(p)}\right] du \leq \int_p^1 \left[\frac{F^{-1}_{(t',\infty)}(u)}{F^{-1}_{(t',\infty)}(p)}\right] du, \text{ for all } p \in (0,1)\,. \tag{17}$$

From (16) we see that (17) is equivalent to

$$\frac{\int_p^1 F^{-1}\left[(1-F(t))\,u+F(t)\right] du}{F^{-1}\left[(1-F(t))\,p+F(t)\right]} \leq$$

$$\frac{\int_p^1 F^{-1}\left[(1-F(t'))\,u+F(t')\right] du}{F^{-1}\left[(1-F(t'))\,p+F(t')\right]},\ 0<t<t'<1,\ p\in(0,1)\,. \tag{18}$$

A change of variable shows that (18) holds if and only if

$$\frac{\int_{(1-F(t))p+F(t)}^1 F^{-1}(x)\,dx}{(1-F(t))\,F^{-1}\left[(1-F(t))\,p+F(t)\right]} \leq \tag{19}$$

$$\frac{\int_{(1-F(t'))p+F(t')}^1 F^{-1}(x)\,dx}{(1-F(t'))\,F^{-1}\left[(1-F(t'))\,p+F(t')\right]},\ 0<t<t'<1,\ p\in(0,1)\,.$$

Substituting $v=(1-F(t))\,p+F(t)$ and $u=(1-F(t'))\,p+F(t')$ we see that (19) is satisfied if and only if

$$\frac{\int_v^1 F^{-1}(x)\,dx}{(1-v)\,F^{-1}(v)} \leq \frac{\int_u^1 F^{-1}(x)\,dx}{(1-u)\,F^{-1}(u)} \text{ for all } 0<v<u<1 \tag{20}$$

or, equivalently, if (15) is decreasing in p. ■

4. Some models with $\varepsilon_{\overline{L}}(p)$ monotone

From the results in previous sections, the monotonicity of the elasticity of the dual Lorenz curve of a population may indicate a possible underestimation of the inequality (as measured by the Lorenz curve) and the feeling of relative deprivation (as measured by the expected proportional shortfall function) as reported by tax returns. The economic literature contains many parametric models for the Lorenz curve (see, for example, the papers by Kakwani and Podder, 1973; Rasche et al., 1980; Gupta, 1984; Aggarwal, 1984; Arnold, 1986; Arnold et al., 1987; Villaseñor and Arnold, 1989; Basmann et al., 1990; Ortega et al., 1991; Ryu and Slottje, 1996; Sarabia, 1997; Sarabia, Castillo and Slottje, 1999, 2001; Sarabia and Pascual, 2002; Rohde, 2009; Wang, Smyth and Ng, 2009; Sarabia et al., 2010 and Sordo, Navarro and Sarabia, 2014). In this section we collect some models such that the dual Lorenz curve has monotone elasticity.

4.1. Power Lorenz curve

The power Lorenz curve is given by $L(p) = p^k$, with $k \geq 1$ and its dual is given by

$$\overline{L}(p) = 1 - (1-p)^k, \, k \geq 1. \tag{21}$$

For $k = 1$, we have $\varepsilon_{\overline{L}}(p) = \varepsilon_L(p) = k$. In order to show that (21) is decreasing for $k > 1$, note that

$$\varepsilon_{\overline{L}}(p) = \frac{kp(1-p)^{k-1}}{1-(1-p)^k}.$$

Differentiating with respect to p, it is not hard to see that

$$\varepsilon'_{\overline{L}}(p) \leq 0 \text{ if and only } 1 - pk \leq (1-p)^k. \tag{22}$$

Now, define the auxiliary function

$$h(p) = (1-p)^k - (1-pk), \;\; p \in [0,1], \, k > 1.$$

It is easy to see that h is increasing on $[0,1]$. Since $h(0) = 0$, it follows that $h(p) \geq 0$ for every $p \in [0,1]$. This implies that $(1-p)^k \geq 1 - pk$ for every p in $[0,1]$ and from (22) it follows that $\varepsilon_{\overline{L}}(p)$ is decreasing.

4.2. Distorted Lorenz curves

Sordo, Navarro and Sarabia (2014) considered a general method of modeling a family of Lorenz curves by distorting a baseline Lorenz curve, L, as follows

$$L_h(p) = h(L(p)), 0 \leq p \leq 1, \tag{23}$$

where h is a convex distortion function (that is, an increasing function from $[0,1]$ to $[0,1]$ such that $h(0) = 0$ and $h(1) = 1$) and showed that a large number of parametric models for the Lorenz curve adopt the form (23). In this section we provide conditions on the distortion h under which the elasticity of $\overline{L}_h$ (the dual of the distorted curve L_h) inherits the monotonicity of the elasticity of $\overline{L}$ (the dual of the initial curve L). In that follows, denote by $\overline{h}(p) = 1 - h(1-p)$, for $0 \leq p \leq 1$ (observe that $h(p)$ is a convex distortion function if and only if $\overline{h}$ is a concave distortion function).

Theorem 4.1. *Let $L(p)$ be a Lorenz curve and let h be a convex distortion function. Let $L_h(p)$ be a distorted Lorenz curve (DLC) of the form (23) and let $\overline{L}_h(p)$ be its dual. Then*

$$\varepsilon_{\overline{L}_h}(p) = \varepsilon_{\overline{L}}(p) \cdot \varepsilon_{\overline{h}}(\overline{L}(p)) \tag{24}$$

Proof. From (4) and (23) we obtain

$$\varepsilon_{\overline{L}_h}(p) = \frac{pL_h'(1-p)}{1-L_h(1-p)} = \frac{p\,h'(L(1-p))\,L'(1-p)}{1-h(L(1-p))} \text{ for } 0 < p < 1.$$

Using that $L(1-p) = 1-\overline{L}(p)$ and rearranging the expression above we obtain

$$\varepsilon_{\overline{L}_h}(p) = \frac{pL'(1-p)}{\overline{L}(p)} \cdot \frac{\overline{L}(p)\,h'(1-\overline{L}(p))}{1-h(1-\overline{L}(p))}$$

which is (24). ■

Corollary 4.2. *If $\overline{L}$ and $\overline{h}$ have increasing (respectively, decreasing) elasticities, then $\overline{L}_h$ has increasing (respectively, decreasing) elasticity.*

Next, we give some examples of families of DLC of the form (23) such that $\varepsilon_{\overline{L}_h}(p)$ is monotone.

4.2.1. The class $L_\delta(p) = 1-[1-L(p)]^\delta$

The dual of the convex distortion function $h(t) = 1-(1-t)^\delta$, $0 < \delta \leq 1$ has constant elasticity. Therefore, if $L(p)$ is a baseline Lorenz curve such that $\varepsilon_{\overline{L}}(p)$ is increasing (respectively, decreasing) it follows from Corollary 4.2. that the dual of a DLC of the form

$$L_\delta(p) = 1-[1-L(p)]^\delta, 0 < \delta \leq 1,$$

has increasing (respectively decreasing) elasticity.

4.2.2. The hierachical class of Sarabia et al. (1999)

Let h be the convex distortion function defined by $h(t) = t^k, k \geq 1$. We know, from Section 4.1, that $\varepsilon_{\overline{h}}(p)$ is decreasing. Given a baseline Lorenz curve $L(p)$, Sarabia et al. (1999) considered a hierachical class of Lorenz curves of the form

$$L_k(p) = [L(p)]^k, k \geq 1$$

If $\varepsilon_{\overline{L}}(p)$ is decreasing, it follows from Corollary 4.2 that the elasticity of $\overline{L}_k(p)$ is decreasing. As a consequence, the elasticity of the curve $\overline{L}_{k,\delta}(p)$, where

$$L_{k,\delta}(p) = \left[1-(1-p)^{\delta}\right]^{k}, k \geq 1, 0 < \delta \leq 1, \tag{25}$$

is also decreasing (the curve (25) is one of the Lorenz curves in the Pareto hierarchy considered by Sarabia et al., 1999).

4.2.3. The class $L_{\theta}(p) = \dfrac{\theta L(p)}{1-(1-\theta)L(p)}$

Let h be the convex distortion function given by $h(t) = \dfrac{\theta t}{1-(1-\theta)t}$, with $0 < \theta \leq 1$. It is easy to prove that the elasticity of $\overline{h}$ is the function

$$\varepsilon_{\overline{h}}(t) = \frac{\theta}{\theta+(1-\theta)t},$$

which is decreasing in the interval $(0,1)$. Therefore, it follows from Corollary 4.2 that if $\overline{L}$ has decreasing elasticity, the family of DLC of the form

$$L_{\theta}(p) = \frac{\theta L(p)}{1-(1-\theta)L(p)}, 0 \leq p \leq 1, 0 < \theta \leq 1,$$

considered by Sordo, Navarro and Sarabia (2014) has decreasing elasticity.

4.2.4. Wang-Smyth-Ng model

Let h be the convex distortion function defined by

$$h_{\beta,\gamma}(t) = 1-(1-t)\exp[-\gamma[1-(1-p)^{1/\beta}]], 0 < \beta \leq 1, \gamma > 0.$$

The elasticity of $\overline{h}_{\beta,\gamma}$, given by

$$\varepsilon_{\overline{h}_{\beta,\gamma}}(t) = 1+\frac{\gamma}{\beta}t^{1/\beta},$$

is increasing in $t \in (0,1)$. From Corollary 4.2, if $\overline{L}$ has an increasing elasticity then

$$h_{\beta,\gamma}(L(p)), 0 < \beta \leq 1, \gamma > 0, \tag{26}$$

has also increasing dual elasticity. The family (26) was considered by Sordo, Navarro and Sarabia (2014). In particular, by taking $L(p) = 1-(1-p)^{\beta}$, we obtain the class of Lorenz curves suggested by Wang, Smyth and Ng (2009). All the curves in this class have increasing elasticity.

5. An illustration using real data

Personal income can be measured using different sources of information. In Europe, the main source is the European Union Survey of Income and Living Conditions (EU-SILC) conducted by the Central Statistics Office. Alternative sources include, among other surveys and administrative data (such as those from Social Security records), data from tax income returns. In Spain, for example, taxation microdata are available under request from the Institute of Fiscal Studies (IFE), an institution attached to the Ministry of the Economy through the State Secretariat for Taxation and Budgets. Although, undoubtedly, EU-SILC and tax income returns taken together complement each other, any analysis of inequality based on the separate interpretation of data from tax returns requires caution because these data exclude people with very low taxable income. If we ignore, for the sake of argument, some issues related to the nature of data[2], the study population (or tax filers) becomes a subset, obtained via lower truncation, of a hypothetical reference population which is the same as the reference population of EU-SILC. A possible underestimation of inequality as reported by tax returns may be anticipated using a simple visual of the elasticity plot of the adjusted dual Lorenz curve for this reference population.

In order to illustrate this issue, we have carried out a descriptive study of the function $\varepsilon_{\overline{L}}(p)$ using data from the EU-SILC 2010 survey, which provides income data of 225,987 households and covers 29 European countries. The variable under study is the "total disposable income of the household", adjusted to take into account that we are dealing with individuals who are members of households of different size and composition (we make this adjustment employing the modified OECD equivalence scale). The unit of analysis chosen is the individual; the income assigned to each individual is the total income of the household to which they belong, adjusted according to the equivalence scale to ensure comparability (see Eurostat, 2010).

Taking in mind the expression (8), we have computed the function $\varepsilon_{\overline{L}}(p)$ from data in the following way. If $x_{(i)}$ denotes the i-th ordered income in the sample of size n, and ω_i denotes its corresponding sample weight[3], for $i = 1,2,\ldots,n-1$, we calculate the points

[2] For example, EU-SILC refers to individuals living in households and tax income returns refer to taxpayers. We are deliberately ignoring that members of the same family or household may file separate tax returns.

[3] Due to the use of sophisticated sampling techniques of stratification, rotation and non-response adjustment, microdata provided by the EU-SILC survey are weighted according to specific sample designs.

$$\left(\frac{\sum_{k=1}^{i} \omega_k}{\sum_{k=1}^{n} \omega_k}, \frac{x_{(j)} \sum_{k=1}^{i} \omega_k}{\left(\sum_{k=1}^{j} \omega_k - \sum_{k=i+1}^{n} \omega_k \right) x_{(j)} + \sum_{k=j+1}^{n} \omega_k x_{(k)}} \right) \tag{27}$$

where j is the index such that $\sum_{k=1}^{j-1} \omega_k < \sum_{k=i+1}^{n} \omega_k \leq \sum_{k=1}^{j} \omega_k$ when $j \geq 2$ and $j = 1$ in case $0 < \sum_{k=i+1}^{n} \omega_k \leq \omega_1$. Observe that this set of points can be considered as an analog estimation of the graph of the elasticity of the dual Lorenz curve associated to the income

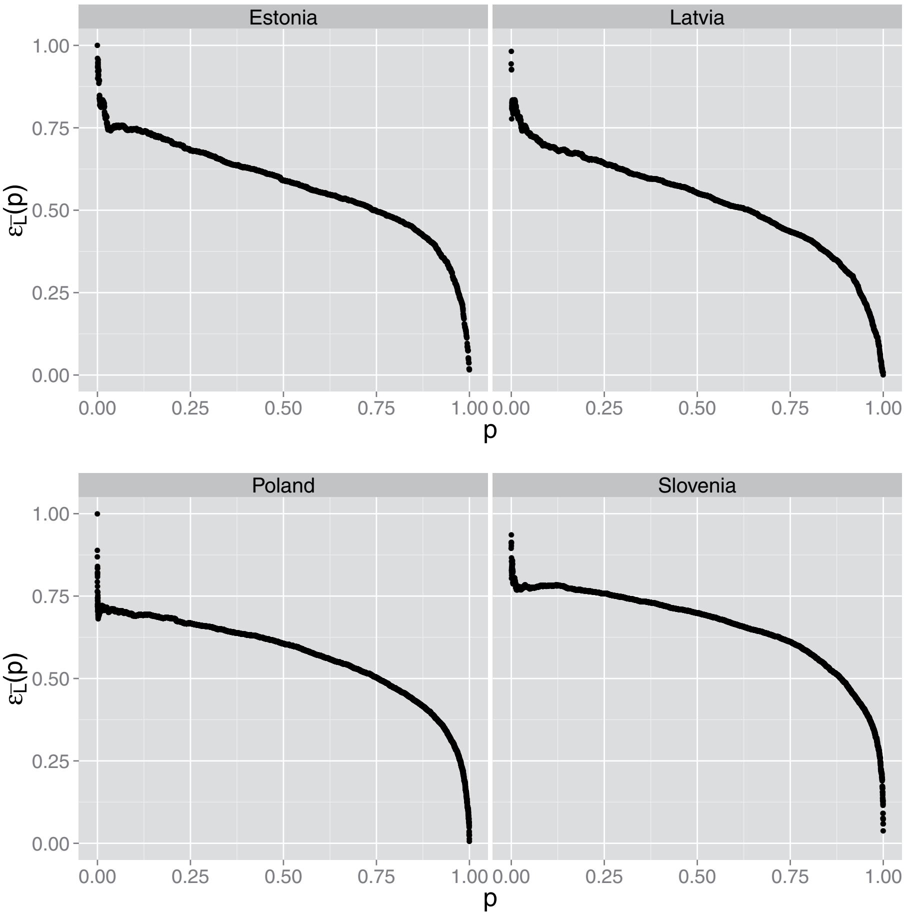


Figure 1: *$\varepsilon_{\overline{L}}(p)$ calculated for Estonia, Latvia, Poland and Slovenia. Source: Generated by authors based on data from EU-SILC 2010.*

distribution. However, since this study is purely illustrative, we have not considered the inferential properties of this estimation (and consequently, we can not discuss about the statistical significance of the results). For the sake of reproducibility, the R code used to calculate the set of points in (27) can be found at Github.[4]

From the results of this study, we conclude that the shapes of the computed elasticity curves can be grouped in basically two different types:

(a) For some countries, the dual Lorenz curve shows a decreasing elasticity. It follows from Corollary 2.2 and Theorem 3.3 that $X \geq_L X_{(t,\infty)}$ and $X \geq_{ep} X_{(t,\infty)}$ for all t, which suggests that statistics from tax returns may under-report inequality and relative deprivation (this is the case of Estonia, Latvia, Poland and Slovenia, see Figure 1).

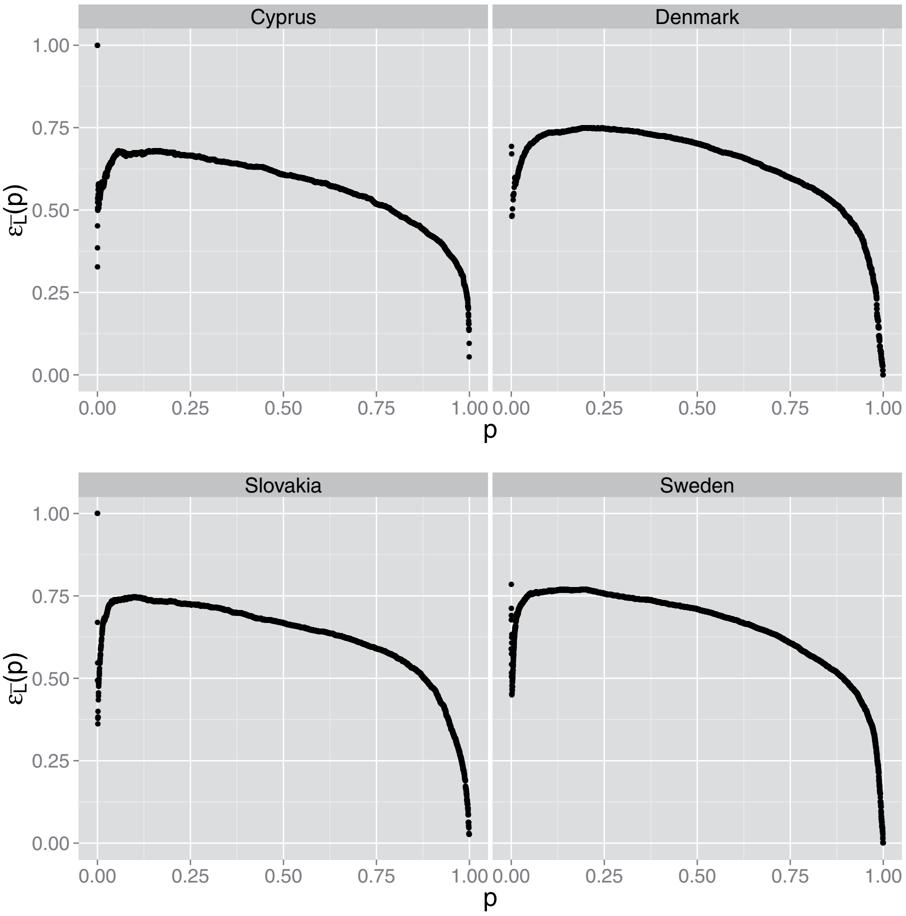


__Figure 2:__ $\varepsilon_{\overline{L}}(p)$ calculated for Cyprus, Denmark, Slovakia and Sweden. Source: Generated by authors based on data from EU-SILC 2010.

[4] https://gist.github.com/AngelBerihuete/fdb11a7dc3ece81bcf5d6261a49af440

(b) For some countries the elasticity curve presents an U inverted shape (this is the case of Cyprus, Denmark, Slovakia and Sweden, see Figure 2). In Denmark, for example, $\varepsilon_{\overline{L}}(p)$ increases in $p \in (0, 0.21)$ and then decreases. From Theorem 2.1 this implies that

$$X_{(t,\infty)} \leq_L X_{(t',\infty)} \text{ for all } t,t' \text{ such that } F^{-1}(0.79) < \ t < t'.$$

Thus, for example, the inequality among the 10% richer of the population is higher than the inequality among the 20% richer. In this case, the elasticity $\varepsilon_{\overline{L}}(p)$ does not provide conclusive information on the relation, in terms of inequality, among tax filers and the entire population.

Acknowledgements

We are grateful to two anonymous referees and the associate editor for their valuable comments, which have improved the presentation and contents of this paper. Miguel A. Sordo and Carmen Ramos acknowledge the support received from Ministerio de Economía y Competitividad (Spain) under grant MTM2014-57559-P.

References

Aggarwal, V. (1984). On optimum aggregation of income distribution data. *Sankhya B*, 46, 343–355.

Arnold, B.C. (1986). A class of hyperbolic Lorenz curves. *Sankhya A*, 48, 427–436.

Arnold, B.C. (1987). Majorization and the Lorenz order: A brief introduction. *Lecture Notes in Statistics,* 43, Springer-Verlag.

Arnold, B.C., Brockett, P.L., Robertson, C.A. and Shu, B.Y. (1987). Generating ordered families of Lorenz curves by strongly unimodal distributions. *Journal of Business and Economic Statistics,* 5, 305–308.

Atkinson, A.B., Piketty, T. and Saez, E. (2011). Top incomes in the long run of history. *Journal of Economic Literature,* 49, 3–71.

Basmann, R.L., Hayes, K.J., Slottje, D.J. and Johnson,J.D. (1990). A general functional form for approximating the Lorenz curve. *Journal of Econometrics*, 43, 77–90.

Belzunce, F., Candel, J. and Ruiz, J. M. (1995). Ordering of truncated distributions through concentration curves. *Sankhya A*, 57, 375–383.

Belzunce, F., Candel, J. and Ruiz, J. M. (1998). Ordering and asymptotic properties of residuals income distributions. *Sankhya B*, 60, 331–348.

Belzunce, F., Pinar, J.F., Ruiz, J.M. and Sordo, M.A. (2012). Comparisons of risks based on the expected proportional shortfall. *Insurance: Mathematics and Economics*, 51, 292–302.

Belzunce, F., Pinar, J.F., Ruiz, J.M. and Sordo, M.A. (2013). Comparison of concentration for several families of income distributions. *Statistics and Probability Letters*, 83, 1036–1045.

Belzunce, F., Martínez-Riquelme, C., Ruiz, J.M. and Sordo, M.A. (2016). On the Comparison of Relative Spacings with Applications. *Methodology and Computing in Applied Probability*. First online. Doi:10.1007/s11009-016-9479-6

Bhattacharya, N. (1963). A property of the Pareto distribution. *Sankhya B*, 25, 195–196.

Eurostat (2010). Description of target variables: cross-sectional and longitudinal. *EU-SILC 065/2010*, Eurostat, Luxembourg.

Greselin, F. and Zitikis, R. (2015). Measuring economic inequality and risk: a unifying approach based on personal gambles, societal preferences and references. *Preprint arXiv:1508.00127.*

Gupta, M.R. (1984). Functional form for estimating the Lorenz curve. *Econometrica*, 52, 1313–1314.

Kakwani, N.C. and Podder, N. (1973). On estimation of Lorenz curves from grouped observations. *International Economic Review*, 14, 278–292.

Moothathu, T.S.K. (1986). A characterization of power distribution through a property of the Lorenz curve. *Sankhya B*, 48, 262–265.

Ortega, P., Fernández, M.A., Ladoux, M. and García, A. (1991). A new functional form for estimating Lorenz curves. *The Review of Income and Wealth*, 37, 447–452.

Piketty, T. and Saez, E. (2003). Income Inequality in the United States: 1913-1998. *Quarterly Journal of Economics*, 118, 1–39.

Ramos, H., Peinado, A. and Ollero, J. (2013). Analysis of inequality in fertility curves fitted by gamma distributions. *SORT*, 37, 233–240.

Rasche, R.H., Gaffney, J., Koo, A. and Obst, N. (1980). Functional forms for estimating the Lorenz curve. *Econometrica*, 48, 1061–1062.

Rohde, N. (2009). An alternative functional form for estimating the Lorenz curve. *Economics Letters*, 105, 61–63.

Ryu, H. and Slottje, D. (1996). Two flexible functional forms for approximating the Lorenz curve. *Journal of Econometrics*, 72, 251–274.

Saez, E. and Zucman, G. (2014). Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data. *NBER Working Paper 20625,* http://eml.berkeley.edu/˜saez/saez-zucmanNBER14wealth.pdf.

Sarabia, J.M. (1997). A hierarchy of Lorenz curves based on the generalized Tukey's lambda distribution. *Econometric Reviews*, 16, 305–320.

Sarabia, J.M., Castillo, E., and Slottje, D.J. (1999). An ordered family of Lorenz curves. *Journal of Econometrics*, 91, 43–60.

Sarabia, J.M., Castillo, E. and Slottje, D.J. (2001). An exponential family of Lorenz curves. *Southern Economic Journal*, 67, 748–756.

Sarabia, J.M. and Pascual, M. (2002). A class of Lorenz curves based on linear exponential loss functions. *Communications in Statistics-Theory and Methods*, 31, 925–942.

Sarabia, J. M., Gómez-Déniz E., Sarabia M. and Prieto, F. (2010). Revisiting a functional form for the Lorenz curve. *Economics Letters*, 107, 249–252.

Sarabia, J. M., Prieto, F. and Trueba, C. (2012). Modeling the probabilistic distribution of the impact factor. *Journal of Infometrics*, 6, 66–79.

Shaked, M. and Shanthikumar, J.G. (2007). *Stochastic orders. Series: Springer Series in Statistics*, Springer.

Sordo, M.A., Navarro, J. and Sarabia, J.M. (2014). Distorted Lorenz curves: models and comparisons. *Social Choice and Welfare*, 42, 761–780.

Stone, C., Trisi, D., Sherman, A. and DeBot, B. (2015). *A Guide to Statistics on Historical Trends in Income Inequality*, Washington, DC: Center on Budget and Policy Priorities.

Villaseñor, J.A. and Arnold, B.C. (1989). Elliptical Lorenz curves. *Journal of Econometrics*, 40, 327–338.

Wang, Z., Smyth, R. and Ng, Y.K. (2009). A new ordered family of curves with an application to measuring income inequality and poverty in rural China. *China Economic Review,* 20, 218–235.

SORT 41 (1) January-June 2017, 73-92 DOI: 10.2436/20.8080.02.51

Comparison of two discrimination indexes in the categorisation of continuous predictors in time-to-event studies

Irantzu Barrio[*,1,5], María Xosé Rodríguez-Álvarez[2,6,7], Luis Meira-Machado[3]
Cristobal Esteban[4,5] and Inmaculada Arostegui[1,5,6]

Abstract

The Cox proportional hazards model is the most widely used survival prediction model for analysing time-to-event data. To measure the discrimination ability of a survival model the concordance probability index is widely used. In this work we studied and compared the performance of two different estimators of the concordance probability when a continuous predictor variable is categorised in a Cox proportional hazards regression model. In particular, we compared the c-index and the concordance probability estimator. We evaluated the empirical performance of both estimators through simulations. To categorise the predictor variable we propose a methodology which considers the maximal discrimination attained for the categorical variable. We applied this methodology to a cohort of patients with chronic obstructive pulmonary disease, in particular, we categorised the predictor variable forced expiratory volume in one second in percentage.

MSC: 62N01, 62P10.

Keywords: Categorisation, prediction models, cutpoint, Cox model.

[*] Corresponding author: E-mail: irantzu.barrio@ehu.eus, Tel.: +34-946012504.
Address: Departamento de Matemática Aplicada, Estadística e Investigación Operativa. Facultad de Ciencia y Tecnología. Universidad del País Vasco UPV/EHU. Barrio Sarriena s/n. 48940 Leioa.

[1] Departamento de Matemática Aplicada, Estadística e Investigación Operativa. Universidad del País Vasco UPV/EHU.

[2] Departamento de Estadística e Investigación Operativa and Biomedical Research Centre (CINBIO). Universidade de Vigo.

[3] Centre of Mathematics and Department of Mathematics and Applications. University of Minho, Portugal.

[4] Servicio de Neumología, Hospital Galdakao-Usansolo.

[5] Red de Investigación en Servicios de Salud en Enfermedades Crónicas (REDISSEC).

[6] BCAM-Basque Center for Applied Mathematics.

[7] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain.

Received: May 2016
Accepted: December 2016

1. Introduction

In the medical field, prediction models have been gaining importance as a support for decision-making, whereby the increased knowledge of potential predictors helps the decision-making process. When the interest relies on predicting the survival time of patients with a certain disease, survival prediction models are commonly used. Particularly, the Cox model (Cox and Oakes, 1984) is the one most widely used in the medical field (Steyerberg et al., 2013).

The development of prediction models may require assumptions about the relationship between the covariates and the response variable. For instance, a common practice in medical research is to categorise continuous predictor variables when a linear relationship does not hold (Turner et al., 2010; Barrio et al., 2016).

The selection of an optimal cutpoint for prognosis purposes has been largely discussed in the literature. For instance, Faraggi and Simon (1996) proposed a cross validation approach to select the cutpoint to classify patients into two risk groups based on the minimisation of the significance level of the logrank test proposed by Lausen and Schumacher (1996). Later, Sima and Gönen (2013) proposed the maximisation of the concordance probability (Gönen and Heller, 2005) as the criterion to dichotomise a continuous predictor. In addition, Liu and Jin (2015) and Rota et al. (2015) have recently proposed non parametric methods to select a time-dependent optimal cutpoint to classify individuals as diseased or disease-free at a given time point t.

However, the aim of this work differs from those presented above. Our goal is to categorise the predictor variable into any possible number of categories to be incorporated in a prediction model, whereas when looking for a unique cutpoint the goal is to classify a patient as diseased or disease-free at a certain time point. This work was motivated in the context of the Stable-COPD study (Esteban et al., 2014) where a model was developed to predict five-year survival in patients with a stable chronic obstructive pulmonary disease (COPD). Clinical researchers aimed to use a categorised version of the predictor variable forced expiratory volume in one second in percentage ($FEV_{1\%}$) in a multiple survival model, but there was no agreement regarding the selection of the optimal cutpoints.

In the context where the outcome of interest takes only two possible values, a proposal has been done to categorise a continuous predictor variable in a logistic regression model by maximising the area under the receiver operating characteristic (ROC) curve (AUC)(Barrio et al., 2016). In fact, the proposal consists of the categorisation of a continuous predictor variable such that the discriminative ability of the prediction model for the categorised variable is maximised. In the context of survival regression models, as far as we know, no proposal has been done to categorise continuous predictor variables. In this paper, we propose to categorise a continuous predictor variable in a Cox proportional hazards regression model as an extension of the work proposed by Barrio et al. (2016) and based in part on the work done for a single cutpoint by Sima and Gönen (2013). However, the challenge is how to measure the discriminative ability of a survival model. Established concepts for binary outcomes have been commonly used by researchers, yet a standard approach has not emerged (Pepe et al., 2008; Schmid and Potapov, 2012). A commonly used parameter is the concordance probability, a global measure which has been defined differently in the literature (Liu and Jin, 2015). In this paper we studied and compared the performance of two different discrimination ability estimators named the c-index (Harrell et al., 1982) and the concordance probability estimator (CPE, Gönen and Heller (2005)) as the parameters to maximise in the categorisation process. Therefore, the goal of this paper is to compare the performance of the CPE and c-index as concordance probability estimators to maximise in the location of optimal cutpoints to categorise continuous predictors in a Cox proportional hazards regression model.

The rest of the paper has been organised as follows. Section 2 outlines the method proposed for categorising continuous variables in a Cox proportional hazards regression model. In Section 3, the performance of the proposed methodology is investigated through simulations. Section 4 provides a description of the Stable-COPD study of stable patients with COPD and the application of the proposed methodology to this data set. Finally, the main conclusions of our paper and some practical recommendations are deferred to Section 5.

2. Methods

This section describes the proposed methodology to categorise a continuous predictor variable in a Cox proportional hazards regression model. We begin by introducing the needed notation and background in Section 2.1 and Section 2.2 and we describe the approach to categorise a continuous predictor variable in Section 2.3.

2.1. Notation and preliminaries

Let T be a non-negative random variable representing the time until the event of interest occurs. As usual, we assume that these event times might be subject to univariate right-censoring denoted by C, which we assume to be independent of T. Let $\mathbf{Z} = [Z_1 \dots Z_p]^{\mathsf{T}}$ be a set of time invariant predictor variables in which we may be interested in terms of studying their relationship with the survival time T.

The most widely used survival regression model is the semiparametric Cox proportional hazards model (Cox, 1972), where the hazard function for T in a time t given the covariate vector $\mathbf{Z}$ is expressed as,

$$h(t|\mathbf{Z}) = h_0(t)\exp(\mathbf{Z}^{\mathsf{T}}\boldsymbol{\beta}) \tag{1}$$

where $h_0(t)$ is the baseline hazard function and $\boldsymbol{\beta}$ is the regression coefficients vector.

2.2. Discriminative ability of a prediction model

In general, it is common to measure the discriminative ability of a prediction model by the concordance probability (Gönen and Heller, 2005). In a setting where the outcome is time-to-event, the concordance probability is usually defined as (Pencina and D'Agostino, 2004)

$$\mathfrak{C} = P(\widetilde{T}_l > \widetilde{T}_m | T_l > T_m), \tag{2}$$

where l and m denote two independent individuals, T_j is the actual survival time of subject j, and $\widetilde{T}_j$ is the predicted survival time provided by the survival prediction model under evaluation.

Under the Cox proportional hazards regression model (1), Pencina and D'Agostino (2004) showed that the concordance probability for the predicted survival times defined in equation (2) is equivalent to the concordance probability for the predicted probability of survival and thus equivalent to the concordance probability defined in terms of the linear predictor of the Cox proportional hazards model given in (1), i.e.,

$$\mathfrak{C} = P(S(t|\mathbf{Z}_l) > S(t|\mathbf{Z}_m)|T_l > T_m) = P(\eta_l < \eta_m | T_l > T_m), \quad \forall t \tag{3}$$

where $S(t|\mathbf{Z}_j) = P(T_j \geq t|\mathbf{Z}_j)$ and $\eta_j = \mathbf{Z}_j^{\mathrm{T}}\boldsymbol{\beta}$. If the concordance probability takes a value of 0.5 then the resulting model provides non informative predictions whereas models predicting better than chance should result in values of $\mathfrak{C}$ lying in the interval $(0.5, 1)$.

From now on, let us denote as $\{\mathbf{z}_i, y_i, \delta_i\}_{i=1}^N$ a sample of size N, where $\mathbf{z}_i$ represents the observed value of the predictor variables for subject i, y_i represents the observed follow-up time for subject i, being the minimum between the censoring (c_i) and the event (t_i) times, i.e. $y_i = min(t_i, c_i)$, and δ_i represents whether subject i is an event ($\delta_i = 1$) or is censored ($\delta_i = 0$). Thus, $\delta_i = I(t_i \leq c_i)$.

In the presence of right censoring, it is difficult to estimate the concordance probability because a problem arises with the comparison of predicted and observed survival times. Harrell et al. (1982) proposed an estimator for the concordance probability called the c-index which is defined as "the proportion of all pairs of patients for which we could determine the ordering of survival times such that the predictions are concordant". More specifically, Harrell et al. (1982) classified the pairs of individuals as *usable* or *unusable*. A pair of individuals is considered unusable in two different situations. One, when both individuals had the event at the same time and, two, if the following time for the individual without the event was shorter than the time until the event for the individual having the event. Thus, the c-index estimator proposed by Harrell et al. (1982) is the proportion of usable individual pairs in which the estimated survival times and the observed survival times are concordant and is computed by forming all pairs of observed data where the individual with the shorter follow-up time is an event. Specifically, the c-index estimator proposed by Harrell et al. (1982) for model (1) would have the following expression

$$c = \frac{\sum\sum_{i<j} \{I(y_i < y_j) I(\widehat{\eta}_i > \widehat{\eta}_j)\delta_i + I(y_j < y_i) I(\widehat{\eta}_j > \widehat{\eta}_i)\delta_j\}}{\sum\sum_{i<j} \{I(y_i < y_j)\delta_i + I(y_j < y_i)\delta_j\}}, \tag{4}$$

where $\widehat{\eta}_j = \mathbf{z}_j^{\mathsf{T}} \widehat{\boldsymbol{\beta}}$, being $\widehat{\boldsymbol{\beta}}$ the vector of the estimated regression coefficients of the Cox proportional hazards regression model.

Even though it is widely used in practice, as pointed out by Gönen and Heller (2005), the c-index estimator proposed by Harrell et al. (1982) is biased and the bias increases with the censoring rate. Hence, Gönen and Heller (2005) proposed an alternative estimator called the concordance probability estimator (CPE), which under the proportional hazards assumption is a consistent estimator of the concordance probability. This estimator is defined as

$$CPE = \frac{2}{N(N-1)} \sum\sum_{i<j} \left\{ \frac{I(\widehat{\eta}_i > \widehat{\eta}_j)}{1 + e^{\widehat{\eta}_j - \widehat{\eta}_i}} + \frac{I(\widehat{\eta}_j > \widehat{\eta}_i)}{1 + e^{\widehat{\eta}_i - \widehat{\eta}_j}} \right\}. \tag{5}$$

Although it has been usually overlooked in the literature, we would like to note that the definition of concordance probability given by Gönen and Heller (2005) (see equation (1) in that paper), differs from that defined on equation (3). In fact, the CPE given in (5) represents an estimator of $P(T_l > T_m | \eta_l < \eta_m)$. Hence the c-index and the CPE estimate, in general, different quantities.

Different estimators have been proposed in the literature to estimate the concordance probability (Schmid and Potapov, 2012). In this paper, we focused on the c-index and the CPE for two main reasons. First, Schmid and Potapov (2012) carried out a comparison of different discrimination indexes and none of the estimators proved to be stable in all scenarios. In addition, previous work has been done on the comparison of these two estimators in the selection of an optimal cutpoint in a Cox proportional hazards regression model and we intended to extend the research done by Sima and Gönen (2013) to the categorisation of a continuous predictor variable in a multiple Cox proportional hazards prediction model.

2.3. Selection of optimal cutpoints in Cox proportional regression models

Let X be a continuous predictor variable which we want to categorise in a Cox proportional hazards regression model considering the presence of other p predictors, $Z_1, \ldots, Z_p$. Our proposal is to categorise X in such a way that the best multiple predictive survival model is obtained, considering the maximal concordance probability achieved. Specifically, given k the number of cutpoints set for categorising X in $k+1$ intervals, let us denote X_{cat_k} the corresponding categorised variable taking values from 0 to k, and $\boldsymbol{x}_k = [x_1 \ldots x_k]^{\mathsf{T}}$ the vector of k cutpoints which maximises the discriminative ability of the Cox proportional hazards regression model in equation (6):

$$h(t|Z_1,\ldots,Z_p,X_{cat_k}) = h_0(t)e^{\sum_{r=1}^{p}\alpha_r Z_r + \sum_{q=1}^{k}\beta_q 1_{\{X_{cat_k}=q\}}}. \tag{6}$$

Note that in this expression the linear predictor η is in fact $\sum_{r=1}^{p}\alpha_r Z_r + \sum_{q=1}^{k}\beta_q 1_{\{X_{cat_k}=q\}}$.

To estimate the vector of the cutpoints of X that maximises the concordance probability, we propose to make use of the algorithms *AddFor* and *Genetic* proposed by Barrio et al. (2016). The former looks sequentially for the k cutpoints whereas the later looks for the vector of the optimal cutpoints using genetic algorithms. This implies that the *Genetic* algorithm is computationally more expensive than the *AddFor*. Nevertheless, it has been proven to perform better specially when two cutpoints are looked for (Barrio et al., 2016). For this reason, in this paper we have limited to the use of the *Genetic* algorithm. In addition, the optimal number of cutpoints can be selected by means of a bootstrap confidence interval for the difference of the bias-corrected concordance probability as proposed by Barrio et al. (2016) and extended here to the Cox proportional hazards setting. Detailed information regarding this approach can be seen in the Supplementary Material.

Note that our approach can be easily applied also to the univariate Cox proportional hazards model in which no other predictors $\mathbf{Z}$ are present. However, in this case there will be many ties on the linear predictor and hence the expressions given in equations (4) and (5) need to be modified accordingly (see Appendix for further details).

3. Simulation study

In this section we present a simulation study conducted to analyse the empirical performance of the methodology proposed in Section 2. We report here the results obtained in this study and compare the performance of the two concordance probability estimators considered. The simulation study is explained in detail below.

All computations were performed in (64 bit) R 3.2.3 (R Core Team, 2016) and a workstation equipped with 24GB of RAM, an Intel Xeon E5620 processor (2.40 Ghz), and Windows 7 operating system. Specifically, the `genoud` function of the `rgenoud` (Mebane and Sekhon, 2011) package was used to compute the genetic algorithms, the `cph` function of the `rms` package (Harrell, 2015) was used for the estimation of the Cox proportional hazards regression model and the c-index, and the `phcpe2` function of the package `CPE` (Mo et al., 2012) was used to estimate the CPE.

3.1. Scenarios and set-up

To simulate the data we assumed that X is a continuous predictor variable normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 2$ and Z a continuous predictor variable normally distributed with mean $\mu = 1$ and standard deviation $\sigma = 1$. Considering the theoretical optimal cutpoints, $c_1, c_2, \ldots, c_k$, we built a categorical variable, X_{cat_k}, such that $X_{cat_k} = 0$ if $X \leq c_1$, $X_{cat_k} = 1$ if $c_1 < X \leq c_2, \ldots$, and $X_{cat_k} = k$ if $X > c_k$. Survival times T were generated assuming a Weibull baseline hazard function such that the Cox proportional hazards model is satisfied (Meira-Machado and Faria, 2014). Specifically,

$$T = \left(\frac{-\,log(1-U)}{\lambda e^{\beta_1 1_{\{X_{cat_k}=1\}} + \ldots + \beta_k 1_{\{X_{cat_k}=k\}} + \alpha Z}} \right)^{1/\gamma}, \tag{7}$$

were U follows a uniform distribution on the interval $(0,1)$, and λ and γ denote the scale and shape parameters of the Weibull distribution respectively. An independent uniform censoring time C was generated, according to the uniform model $U(0,\tau)$, and the event indicator δ was defined as $I(T \leq C)$. The parameter τ was chosen to obtain censoring percentages of about 20%, 50% and 70%. Simulations were performed for sample sizes of $N = 500$ and $N = 1000$. In all cases, $R = 500$ replicates of simulated data were performed.

Several settings were considered in this simulation study, which are summarised in Table 1. First of all, we considered $k = 1, 2$ and 3 as the number of cutpoints. In Scenario I, $k = 1$ was considered with three different alternatives for the theoretical cutpoint a) centred on the distribution of X, i.e., $c_1 = 0$; b) shifted to the high risk area, $c_1 = 1.5$; and c) shifted to the low risk area, $c_1 = -1.5$. In Scenario II we considered two theoretical cutpoints $c_1 = -1$ and $c_2 = 1$. Finally three theoretical cutpoints $c_1 = -1.5, c_2 = 0$ and $c_3 = 1.5$ where considered in Scenario III. In the later scenario, we also considered two different settings, one in which a monotonic increase risk relationship was considered (IIIa) and the other for a non-monotonic risk relationship (IIIb).

The performance of each of the concordance probability estimators considered was evaluated by means of the bias and mean square error (MSE) of the estimated optimal cutpoints for each iteration as follows:

$$MSE_s = \frac{1}{k} \sum_{d=1}^{k} (x_{sd} - c_d)^2$$

where x_{sd} is the estimated d^{th} optimal cutpoint in the simulation s and c_d is the d^{th} theoretical cutpoint.

3.2. Results

Given the large number of proposed scenarios and different conclusions obtained, we begin by summarising the main findings.

Simulation results suggest that, in general, both indexes performed similarly in terms of the mean square error when it comes to low censoring rates (20%). However, for large censoring rates (70%), the c-index performed better than the CPE in all scenarios considered. As could have been expected, in all cases the bias and MSE decrease as the sample size increase.

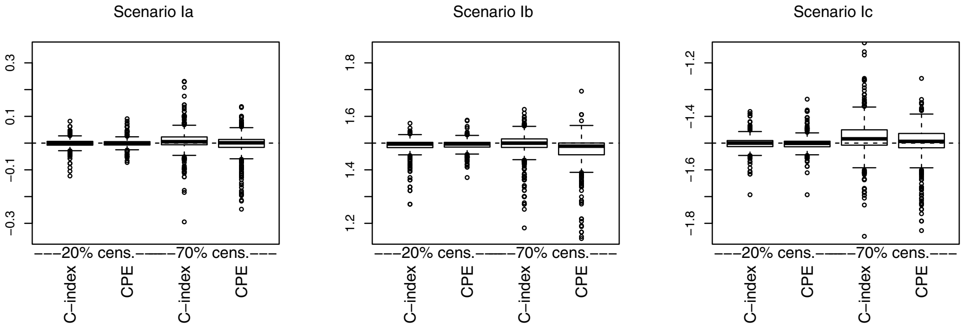


Figure 1: *Boxplot of the estimated optimal cutpoints based on 500 simulated data sets,* $N = 500$ *sample size and one theoretical cutpoint. Results are shown for censoring rates of* 20% *and* 70% *and c-index and CPE discriminative ability estimators. From left to right: (a) theoretical cutpoint,* $c = 0$*; (b) theoretical cutpoint,* $c = 1.5$*; and (c) theoretical cutpoint,* $c = -1.5$*. The theoretical cutpoint is represented with a dashed line.*

Table 1: *Description of the different scenarios considered for the simulation study. Weibull baseline hazard function with shape* γ *and scale* λ*. Uniform censoring* $C\,U(0,\tau)$*.*

Scenario	**Theoretical cutpoints**	**Parameters**	**Censorship (τ)**		
			20%	**50%**	**70%**
Ia	0	$\gamma = 1, \lambda = 0.1$ $\beta_1 = 2.5, \alpha = 1$	11	1.6	0.5
Ib	1.5	$\gamma = 1, \lambda = 0.1$ $\beta_1 = 2.5, \alpha = 1$	19	3.5	1.15
Ic	-1.5	$\gamma = 1, \lambda = 0.1$ $\beta_1 = 2.5, \alpha = 1$	4.75	0.75	0.27
II	-1 & 1	$\gamma = 1, \lambda = 0.1$ $\beta_1 = 1.5, \beta_2 = 2.5$ $\alpha = 1$	8.5	1.5	0.5
IIIa	-1.5 & 0 & 1.5	$\gamma = 1, \lambda = 0.1$ $\beta_1 = 1.5, \beta_2 = 2.5$ $\beta_3 = 3.5, \alpha = 1$	5.25	0.85	0.27
IIIb	-1.5 & 0 & 1.5	$\gamma = 1, \lambda = 0.1$ $\beta_1 = 1.5, \beta_2 = -1$ $\beta_3 = 1.5, \alpha = 1$	21	3.5	1.15

Let us turn now to a more detailed discussion of the results of this study. Figure 1 depicts the boxplot of the estimated optimal cutpoints over 500 simulated data sets, for the c-index and CPE estimators and a sample size of $N = 500$ and censoring rates of 20% and 70% for Scenarios Ia, Ib and Ic, where a single optimal cutpoint is searched for.

Table 2: *Simulations results when one theoretical optimal cutpoints was sought for (Scenarios Ia, Ib and Ic), censorship of 20%, 50% and 70% and the* Genetic *algorithm. Mean, standard deviation, median, bias and mean MSE for the estimated cutpoints are reported when CPE or c-index concordance probability estimators are used as the maximisation criteria.*

Sample size	Cens.	theoretical cutpoint	Cutpoint Estimation							
			CPE				c-index			
			Mean (sd)	Median	Bias	MSE	Mean (sd)	Median	Bias	MSE
			Scenario Ia							
$N = 500$	20%	0	0.000 (0.018)	0.000	0.000	0.000	−0.001 (0.019)	0.000	−0.001	0.000
	50%	0	−0.001 (0.027)	0.000	−0.001	0.001	0.000 (0.025)	0.001	0.000	0.001
	70%	0	−0.007 (0.047)	0.001	−0.007	0.002	0.008 (0.046)	0.005	0.008	0.002
$N = 1000$	20%	0	−0.002 (0.009)	0.000	−0.002	0.000	−0.001 (0.010)	0.000	−0.001	0.000
	50%	0	−0.003 (0.012)	0.000	−0.003	0.000	−0.001 (0.015)	0.000	−0.001	0.000
	70%	0	−0.006 (0.031)	0.000	−0.006	0.001	0.003 (0.026)	0.003	0.003	0.001
			Scenario Ib							
$N = 500$	20%	1.5	1.493 (0.021)	1.497	−0.007	0.000	1.489 (0.032)	1.497	−0.011	0.001
	50%	1.5	1.490 (0.030)	1.495	−0.010	0.001	1.490 (0.037)	1.498	−0.010	0.001
	70%	1.5	1.470 (0.061)	1.488	−0.030	0.005	1.491 (0.051)	1.500	−0.009	0.003
$N = 1000$	20%	1.5	1.498 (0.013)	1.499	−0.002	0.000	1.496 (0.016)	1.499	−0.004	0.000
	50%	1.5	1.496 (0.015)	1.499	−0.004	0.000	1.497 (0.018)	1.500	−0.003	0.000
	70%	1.5	1.483 (0.031)	1.492	−0.017	0.001	1.496 (0.023)	1.500	−0.004	0.001
			Scenario Ic							
$N = 500$	20%	−1.5	−1.501 (0.028)	−1.501	−0.001	0.001	−1.502 (0.029)	−1.500	−0.002	0.001
	50%	−1.5	−1.500 (0.042)	−1.498	0.000	0.002	−1.491 (0.053)	−1.494	0.009	0.003
	70%	−1.5	−1.508 (0.096)	−1.494	−0.008	0.009	−1.478 (0.087)	−1.484	0.022	0.008
$N = 1000$	20%	−1.5	−1.499 (0.015)	−1.500	0.001	0.000	−1.500 (0.015)	−1.500	0.000	0.000
	50%	−1.5	−1.498 (0.021)	−1.498	0.002	0.000	−1.496 (0.025)	−1.498	0.004	0.001
	70%	−1.5	−1.498 (0.047)	−1.495	0.002	0.002	−1.488 (0.048)	−1.492	0.012	0.002

Table 3: *Simulations results when two theoretical optimal cutpoints were sought for (Scenario II), censorship of 20%, 50% and 70% and the* Genetic *algorithm. Mean, standard deviation, median, bias and mean MSE for the estimated cutpoints are reported when CPE or c-index concordance probability estimators are used as the maximisation criteria.*

Sample size	**Cens.**	**theoretical cutpoint**	**Cutpoint Estimation**							
			CPE				**c-index**			
			Mean (sd)	**Median**	**Bias**	**MSE**	**Mean (sd)**	**Median**	**Bias**	**MSE**
			Scenario II							
N=500	20%	−1	−1.007 (0.049)	−1.002	−0.007	0.006	−1.006 (0.059)	−1.000	−0.006	0.011
		1	0.984 (0.098)	0.992	−0.016		0.979 (0.132)	0.996	−0.021	
	50%	−1	−1.010 (0.096)	−1.000	−0.010	0.017	−1.000 (0.118)	−0.993	0.000	0.028
		1	0.969 (0.155)	0.989	−0.031		0.969 (0.203)	0.999	−0.031	
	70%	−1	−1.200 (0.463)	−1.021	−0.200	0.440	−0.997 (0.269)	−0.983	0.003	0.085
		1	0.677 (0.723)	0.957	−0.323		0.998 (0.313)	1.005	−0.002	
N=1000	20%	−1	−1.002 (0.023)	−1.001	−0.002	0.001	−1.002 (0.026)	−0.999	−0.002	0.002
		1	0.991 (0.040)	0.995	−0.009		0.994 (0.062)	0.997	−0.006	
	50%	−1	−1.003 (0.046)	−0.999	−0.003	0.003	−1.004 (0.054)	−0.997	−0.004	0.004
		1	0.987 (0.059)	0.994	−0.013		0.991 (0.075)	0.998	−0.009	
	70%	−1	−1.030 (0.157)	−0.998	−0.030	0.059	−0.997 (0.104)	−0.993	0.003	0.015
		1	0.928 (0.296)	0.982	−0.072		1.000 (0.138)	1.000	0.000	

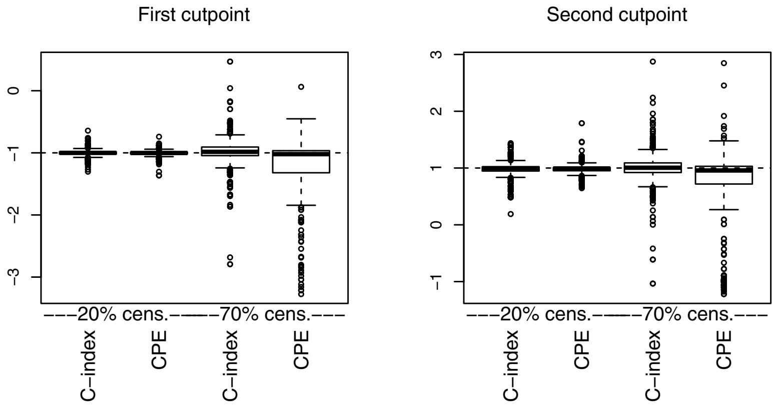


Figure 2: *Boxplot of the estimated optimal cutpoints based on 500 simulated data sets, N* $= 500$ *sample size and two theoretical cutpoints. Results are shown for censoring rates of* 20% *and* 70% *and c-index and CPE discriminative ability estimators. The theoretical cutpoint is represented with a dashed line.*

Numerical results for these scenarios are given in Table 2. As can be seen, our approach performed satisfactorily regardless of the location of the theoretical cutpoint, with, as said before, the c-index performing slightly better for high censoring rates. However, this can not be considered a general rule. Simulations studies conducted in a univariate setting showed that neither the CPE nor the c-index performed satisfactorily, especially when the optimal cutpoint is non centred. These results are presented and discussed in detail in the Supplementary Material (Table B1 and Figure B1).

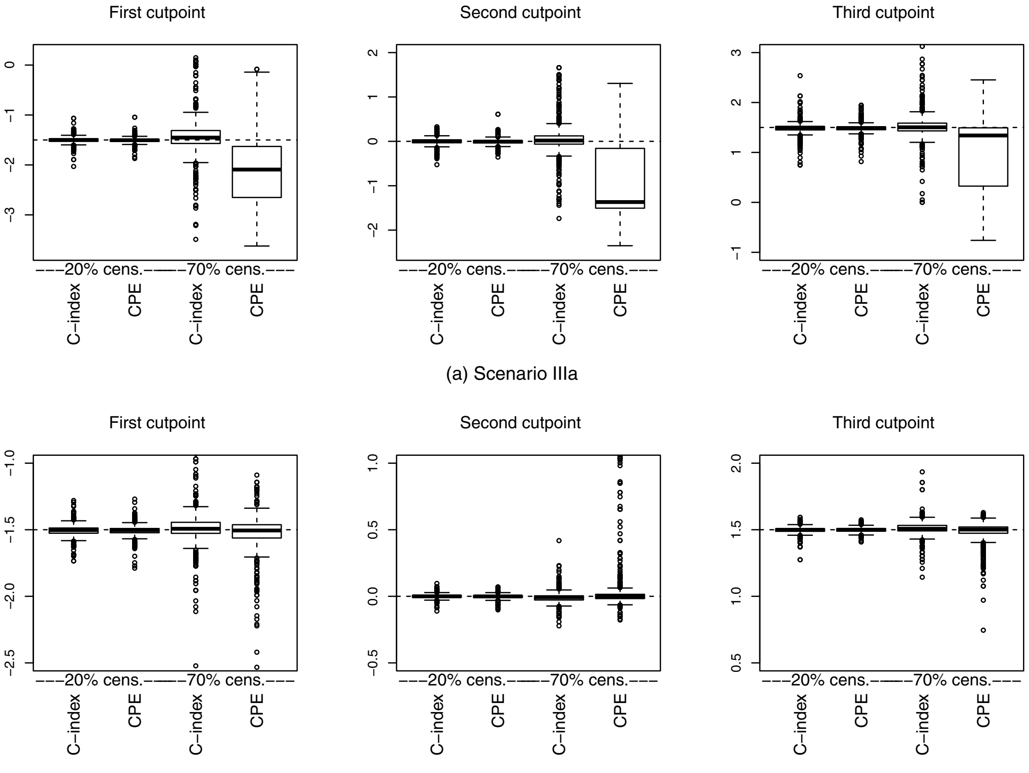
Figure 2 depicts the boxplots of the estimated optimal cutpoints for Scenario II, where two optimal cutpoints are sought for. Numerical results are reported in Table 3. Once again, the c-index outperformed the CPE when high censoring rates were considered. Nevertheless, for censoring rates below 50% both estimators performed satisfactorily.

Finally, Figure 3 depicts the boxplots of the estimated optimal cutpoints for Scenarios IIIa and IIIb, where three optimal cutpoints are sought for a monotonic increasing and non-monotonic risk relationship, respectively. Numerical results are reported in Table 4. These results suggest that the method performed satisfactorily regardless of the risk relationship considered. Nevertheless, for high censoring rates, the CPE performed better when a non-monotonic risk relationship was considered.

We must note that when more than one cutpoint is searched for, the estimated cutpoints have been ordered from the smallest to the largest to classify them as “first”, “second” or “third” cutpoints. This may cause an incorrect classification whenever the estimated smallest cutpoint corresponds to the theoretical “second” cutpoint for example.

Table 4: *Simulations results when three theoretical optimal cutpoints were looked for with a monotonic increasing and non-monotonic relationship with the outcome (Scenarios IIIa and IIIb) and censorship of 20%, 50% and 70%. Mean, standard deviation, median, bias and mean squared error (MSE) for the estimated cutpoints over 500 simulated data sets are reported when CPE or C-index concordance probability estimators are used as the maximisation criteria.*

Sample size	Cens.	theoretical cutpoint	Cutpoint Estimation							
			CPE				c-index			
			Mean (sd)	Median	Bias	MSE	Mean (sd)	Median	Bias	MSE
						Scenario IIIa				
$N = 500$	20%	−1,5	−1.507 (0.066)	−1.505	−0.007	0.008	−1.501 (0.078)	−1.499	−0.001	0.012
		0	−0.006 (0.087)	−0.006	−0.006		−0.002 (0.096)	0.002	−0.002	
		1,5	1.483 (0.107)	1.489	−0.017		1.484 (0.141)	1.495	−0.016	
	50%	−1,5	−1.568 (0.269)	−1.501	−0.068	0.075	−1.486 (0.177)	−1.488	0.014	0.032
		0	−0.078 (0.313)	−0.007	−0.078		−0.006 (0.174)	0.002	−0.006	
		1,5	1.458 (0.211)	1.483	−0.042		1.490 (0.181)	1.495	−0.010	
	70%	−1,5	−2.119 (0.675)	−2.093	−0.619	0.983	−1.442 (0.476)	−1.453	0.058	0.175
		0	−0.951 (0.744)	−1.368	−0.951		0.045 (0.450)	0.018	0.045	
		1,5	1.006 (0.641)	1.340	−0.494		1.523 (0.305)	1.504	0.023	
$N = 1000$	20%	−1,5	−1.499 (0.029)	−1.500	0.001	0.002	−1.499 (0.036)	−1.500	0.001	0.002
		0	−0.003 (0.040)	−0.003	−0.003		−0.004 (0.045)	−0.001	−0.004	
		1,5	1.492 (0.056)	1.498	−0.008		1.487 (0.058)	1.497	−0.013	
	50%	−1,5	−1.500 (0.072)	−1.497	0.000	0.005	−1.490 (0.084)	−1.494	0.010	0.006
		0	−0.005 (0.064)	−0.003	−0.005		0.002 (0.072)	0.001	0.002	
		1,5	1.487 (0.070)	1.496	−0.013		1.491 (0.069)	1.499	−0.009	
	70%	−1,5	−1.739 (0.469)	−1.525	−0.239	0.340	−1.455 (0.211)	−1.479	0.045	0.035
		0	−0.354 (0.635)	−0.025	−0.354		0.027 (0.188)	0.008	0.027	
		1,5	1.335 (0.436)	1.486	−0.165		1.512 (0.154)	1.500	0.012	
						Scenario IIIb				
$N = 500$	20%	−1,5	−1.506 (0.047)	−1.503	−0.006	0.001	−1.508 (0.052)	−1.503	−0.008	0.001
		0	−0.002 (0.019)	−0.001	−0.002		−0.001 (0.018)	0.000	−0.001	
		1,5	1.499 (0.021)	1.500	−0.001		1.496 (0.029)	1.499	−0.004	
	50%	−1,5	−1.513 (0.070)	−1.505	−0.013	0.002	−1.506 (0.086)	−1.500	−0.006	0.003
		0	−0.002 (0.025)	−0.001	−0.002		−0.005 (0.027)	−0.002	−0.005	
		1,5	1.497 (0.033)	1.500	−0.003		1.499 (0.043)	1.502	−0.001	
	70%	−1,5	−1.469 (0.410)	−1.506	0.031	0.073	−1.495 (0.176)	−1.493	0.005	0.013
		0	0.050 (0.198)	−0.001	0.050		−0.010 (0.054)	−0.007	−0.010	
		1,5	1.481 (0.087)	1.501	−0.019		1.508 (0.067)	1.507	0.008	
$N = 1000$	20%	−1,5	−1.504 (0.023)	−1.502	−0.004	0.000	−1.502 (0.024)	−1.501	−0.002	0.000
		0	0.000 (0.010)	0.000	0.000		0.001 (0.010)	0.000	0.001	
		1,5	1.501 (0.012)	1.501	0.001		1.499 (0.013)	1.500	−0.001	
	50%	−1,5	−1.504 (0.034)	−1.502	−0.004	0.001	−1.501 (0.039)	−1.500	−0.001	0.001
		0	0.002 (0.013)	0.000	0.002		0.000 (0.016)	0.000	0.000	
		1,5	1.499 (0.020)	1.500	−0.001		1.500 (0.019)	1.501	0.000	
	70%	−1,5	−1.514 (0.071)	−1.502	−0.014	0.003	−1.504 (0.080)	−1.498	−0.004	0.003
		0	0.003 (0.033)	0.000	0.003		−0.003 (0.023)	−0.003	−0.003	
		1,5	1.494 (0.039)	1.501	−0.006		1.500 (0.036)	1.503	0.000	



(a) Scenario IIIa

(b) Scenario IIIb

Figure 3: *Boxplot of the estimated optimal cutpoints based on 500 simulated data sets,* $N = 500$ *sample size, three theoretical cutpoints, monotonic increasing and non-monotonic relationship. Results are shown for censoring rates of* 20% *and* 70% *and c-index and CPE discriminative ability estimators. From top to bottom: (a) Monotonic increasing relationship (Scenario IIIa); (b) non-monotonic relationship (Scenario IIIb). Theoretical cutpoints are represented with a dashed line.*

4. Application to the Stable-COPD study

Chronic obstructive pulmonary disease (COPD) is one of the most common chronic diseases, its prevalence is expected to increase over the next few decades (Buist et al., 2008), and is a leading cause of death in developed countries. Patients being treated for COPD at five outpatient respiratory clinics affiliated with the Hospital Galdakao-Usansolo in Biscay were recruited in the Stable-COPD study (Esteban et al., 2014). Patients were consecutively included in the study if they had been diagnosed with COPD for at least six months and had been receiving medical care at one of the hospital respiratory outpatient facilities for at least six months. Their COPD had to be stable for six weeks before enrolment. Patients were followed for up to five years. In total, information for 543 patients was obtained of which the 96.13% were men, the mean age was of 68.32 and the 30.76% died in a 5-years period for which the mean survival time was of 2.77 years. The main selected variables collected in this study included sociodemographic variables, forced expiratory volume in one second in percentage ($FEV_{1\%}$), body mass index (BMI), dyspnea measured with the modified scale of the Medical Research Council (mMRC, Fletcher et al. (1959)) and the walking distance among others. A brief description of the main variables used in this paper is given in Table 5. One of the main goals of this study was to develop prediction models for patients with stable COPD.

Table 5: *A description of the selected variables from the Stable-COPD study ($N = 543$).*

Variable	Mean (sd)	Range
Age	68.32 (8.32)	33 - 86
FEV$_{1\%}$	55 (13.31)	18 - 105
BMI	28.28 (4.43)	16.38 - 44.04
Time until event (days)	1574.89 (483.43)	23 - 2045
Walking distance	408.89 (92.43)	46 - 644
Sex‡ – *Men*	522 (96.13 %)	
Dyspnoea‡		
1	69 (12.71 %)	
2	264 (48.62 %)	
3	166 (30.57 %)	
4	23 (4.24 %)	
5	21 (3.87 %)	
5-year mortality‡ – *Yes*	167 (30.76 %)	

‡Categorical variables are shown as absolute and relative frequencies

Table 6: *Airflow obstruction level measured by $FEV_{1\%}$ based on the different cutpoints used in the literature to categorise the continuous $FEV_{1\%}$ variable.*

Criteria	Mild	Moderate	Severe	Very Severe
GOLD	≥ 80	$[50-80)$	$[30-50)$	< 30
BODE	≥ 65	$[50-65)$	$(35-50)$	≤ 35
HADO	> 65	$[50-65]$	$[35-50)$	< 35
ADO	≥ 65		$(35-65)$	≤ 35
DOSE	≥ 50		$[30-50)$	< 30
COCOMICS	≥ 70	$(55-70)$	$(35-55]$	≤ 35

An important predictor for COPD mortality or hospitalisation is $FEV_{1\%}$, which is commonly used by clinicians to diagnose and measure the severity of the disease (Vestbo et al., 2013). Recently, several scores have been proposed which include a categorised version of $FEV_{1\%}$ among the predictor variables. The most commonly used scores are the original BODE index (Celli et al., 2004), HADO index (Esteban et al., 2006), ADO index (Puhan et al., 2009), and DOSE (Jones et al., 2009). Although all prediction scores are based on prediction models which use a categorised version of the predictor variable $FEV_{1\%}$, not all of them use the same cutpoints (see Table 6). To date, the most widely-used cutpoints are the ones proposed by the Global Obstructive Lung Disease (GOLD) guidelines (mild ≥ 80, moderate 50-79, severe 30-49 and very severe < 30, Rabe et al. (2007)). More recently, Almagro et al. (2014) have proposed a new categorisation of $FEV_{1\%}$ to predict five-year survival in COPD patients. This research was framed within the Collaborative Cohorts to Assess Multicomponent Indices of COPD in Spain (COCOMICS) study.

Hence, and taking all this into account, three factors motivated us to look for the best categorisation of the variable $FEV_{1\%}$ as part of the development of the prediction survival model in the Stable-COPD study. First of all, this variable is an important predictor of survival in COPD patients. Since other prediction models and especially clinical guidelines use a categorised version of this variable, the clinicians involved in the study considered it was necessary to include a categorised version of this variable in the prediction model. Second, recent research shows the importance of seeking optimal cutpoints for this variable (Almagro et al., 2014). Third, as indicated above, to date there are no unified criteria on how to categorise the variable $FEV_{1\%}$.

We looked for the best categorisation of the predictor variable $FEV_{1\%}$ in a multivariate setting, taking into account the effect of age and dyspnoea, which are seen as important predictors for the severity of patients with stable COPD (Bestall et al., 1999). In fact, these variables together with a categorisation of $FEV_{1\%}$ are the ones used in the ADO index (Puhan et al., 2009), which turned out to be the best multivariate score to predict 5-year mortality based on the c-index (Marin et al., 2013). The censoring rate in our data set was 66.6%. Considering the results obtained in the simulation study, the c-index concordance probability estimator was used to select the optimal cutpoints since it appeared to perform better under this scenario. To select the optimal number of cutpoints we considered the bootstrap confidence interval for the bias-corrected c-index using $B = 200$ bootstrap replicates. In this data set, the proportional hazards assumption was verified (Grambsch and Therneau, 1994).

In a first stage we looked for $k = 3$ cutpoints and compared them with $k = 2$ cutpoints, which are also the number of cutpoints used in the categorisation of $FEV_{1\%}$ in the ADO index. Using the c-index estimator and the *Genetic* algorithm we obtained that the optimal cutpoints were $(29.32, 50.69)$ and $(29.90, 49.95, 50.54)$ when we looked for $k = 2$ and $k = 3$ number of cutpoints, respectively. In this case, the optimal cutpoints obtained when the CPE was used as the concordance probability estimator were almost the same, being $(29.79, 50.63)$ for $k = 2$ and $(29.69, 49.37, 50.82)$ for $k = 3$. When we compared $k = 2$ versus $k = 3$ number of cutpoints, we obtained a 95% bootstrap CI $(-0.005, 0.015)$ for the difference bias-corrected c-index. Consequently, the optimal number of cutpoints considering the multivariate setting would be $k = 2$, resulting in mild-moderate ($> 50\%$), severe ($[30\% - 50\%]$) and very severe ($< 30\%$) categories. Note that the estimated cutpoints matched up with those used in the DOSE index (Jones et al., 2009) and those proposed in the GOLD guidelines (Rabe et al., 2007). The estimated cutpoint which separated the categories severe from very severe, differed slightly from the one used in the BODE, HADO and ADO indexes i.e., 35, which was identified by the American Thoracic Society (Celli et al., 2004). For illustration purposes, we would like to indicate that the bias corrected c-index for the ADO categorisation proposal was 0.701, lower than the 0.717 obtained using our approach.

5. Discussion

Categorisation of a continuous predictor variable is a commonly used strategy in biomedical research (Turner et al., 2010), where decisions are usually based on the risk classification of patients. To the best of our knowledge, up to now, no approaches have been proposed in the literature for the categorisation of a continuous predictor variable in a multiple Cox proportional hazards regression model. In this paper, we have proposed and validated by means of simulations a methodology to categorise a continuous predictor variable by maximising the concordance probability of the final model for the categorised variable.

Although the objective is different, several methods have been proposed in the literature to select optimal cutpoints (a unique cutpoint) for the prognosis of a disease (Faraggi and Simon, 1996; Sima and Gönen, 2013). In that context, the aim is to select the best cutpoint to dichotomise a variable and classify individuals as diseased or disease-free based on that cutpoint. Sima and Gönen (2013) proposed the maximal discrimination as a method to dichotomise a continuous predictor. They compared the maximisation of the discrimination indexes CPE and c-index together with the maximisation of the log-rank, Wald and partial likelihood ratio statistics for the location of one optimal cutpoint.

Our proposal is different to Sima and Gönen's proposal in one main aspect. Our goal is to categorise a continuous predictor variable to be used in a Cox proportional hazards regression model, considering any possible number of cutpoints. In fact, the most common scores used to predict mortality in COPD patients, such as BODE or ADO, use categorised versions (with more than two categories) of continuous predictors (Celli et al., 2004; Puhan et al., 2009). Furthermore, the methodology that we propose considers the effect that other predictor or confounding variables may have on the selection of the optimal cutpoints. Finally, our proposal allows to select the optimal number of cutpoints to categorise the predictor variable using a bootstrap confidence interval for the difference of the bias-corrected concordance probability estimators.

This proposal is an extension of the methodology proposed by Barrio et al. (2016) for the logistic regression setting. However, in time-to-event studies different estimators of the concordance probability have been proposed. In this paper we have studied and compared the performance of two estimators: the c-index and the CPE, in order to evaluate their performance in the categorisation of a continuous predictor variable in a Cox proportional hazards regression model.

The finite sample performance of the concordance probability estimators considered, i.e., c-index and CPE, was investigated through simulations. Results indicate that both concordance probability estimators performed satisfactorily in a multiple Cox regression model for any number of cutpoints and low-moderate censoring rates ($\leq 50\%$). When the censoring rate considered was high ($> 70\%$), the c-index appeared to outperform the CPE in all the scenarios considered. Additionally, the simulation results for three cutpoints showed that optimal cutpoints can be obtained regardless of the relationship of the latent continuous variable and the outcome. However, when we looked for a unique cutpoint in a univariate Cox proportional hazards regression model, results differed depending on the location of the theoretical cutpoint (results shown in the Supplementary Material). In fact, we observe that for a unique and not centred cutpoint, neither CPE nor c-index performed satisfactorily. Depending on whether the theoretical cutpoint was positively or negatively migrated from the centre of the distribution, smaller bias and MSE values were obtained for CPE or c-index. We must take into a account that when a univariate model is considered and the predictor variable takes only two possible values, there are many ties on the estimated survival probabilities and hence it may have an impact on the estimated concordance probability. Consequently, based on the simulation results obtained, we give the following recommendations for use in practice. For low-moderate censoring rates ($\leq 50\%$), either the c-index or the CPE can be used as maximisation criteria to obtain optimal cutpoints. However, for high censoring rates we recommend the c-index as the concordance probability estimator to maximise. Finally, we do not recommend the use of this proposal for dichotomisation in a univariate model.

Although we tried to evaluate many different scenarios, we could not address all possible real world settings and hence the conclusions we got can be extended only to those situations that were defined in the simulation study. In the scenarios we simulated we considered true optimal cutpoints in order to be able to compare the estimated cutpoints with those theoretical ones. Nevertheless, in practice neither the location or the number of cutpoints are known. We are aware that in theory the optimal number of cutpoints for the categorisation of a continuous variable does not exist, since above all the possible number of cutpoints, the best option would be the continuous variable. However, in clinical practice categorical versions of the continuous variables can be preferred without it always being clear which is the best number of categories to be used. For those situations we provided a proposal to decide among different number of cutpoints based on the bootstrap confidence interval (Barrio et al., 2016) which has been extended to the Cox proportional hazards regression model (see Supplementary Material). Although further research is needed to provide accurate methods for the selection of the optimal number of cutpoints, the results suggest that, when using the c-index, the number of cutpoints can be selected based on the bootstrap CI for the difference of the bias corrected estimated concordance probability.

In this paper we have not considered time-dependent discriminative ability measures as a parameter for selecting optimal cutpoints. Note that the concordance probability index is a global measure that does not take into account the time at which the prediction of the event is desired. This implies that the optimal cutpoints are considered to be the same whatever the time of interest is. However, this may not be necessarily true. To overcome this problem, we are currently working on the application of time-dependent discrimination measures (Heagerty and Zheng, 2005; Pepe et al., 2008) in the search for time-dependent optimal cutpoints.

When we applied the proposed methodology to the Stable-COPD study, we saw that the cutpoints obtained to categorise the predictor variable $FEV_{1\%}$ corresponded to cutpoints previously used in the literature, obtaining clinically valid optimal cutpoints.

To summarise, we have compared the performance of two concordance probability estimators as the maximisation criteria to obtain optimal cutpoints to categorise continuous predictor variables in a Cox proportional hazards regression model. By means of simulations we have seen that the methodology proposed for categorising continuous predictors in a Cox proportional hazards regression model provides the optimal location and number of the cutpoints. Additionally, we have implemented this methodology into an R function which leads to easy use of this methodology in practice.

Funding

This study was supported by grants IT620-13 from the Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco, MTM2013-40941-P, MTM2014-55966-P and MTM2016-74931-P from the Ministerio de Economía y Competitividad and FEDER and the Agrupamento INBIOMED from DXPCTSUG-FEDER unha maneira de facer Europa (2012/273). María Xosé Rodríguez-Álvarez acknowledges financial support for Severo Ochoa Program SEV-2013-0323 and Basque Government BERC Program 2014-2017. Luís Meira-Machado acknowledges financial support for Portuguese Funds through FCT-“Fundação para a Ciência e a Tecnologia”, within Project UID/MAT/00013/2013. The collection of the COPD data used for this study was supported in part by grants PI020510 from the Fondo de Investigación Sanitaria and 200111002, 2005111008, from the Departamento de Salud del Gobierno Vasco and the Research Committee Hospital Galdakao-Usansolo.

Acknowledgement

We would like to thank the referees for their constructive comments and suggestions which led to substantial improvement of this paper.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

Almagro, P., Martinez-Camblor, P., Soriano, J., Marin, J., Alfageme, I., Casanova, C., Esteban, C., Soler-Cataluña, J., De-Torres, J., and Celli, B. (2014). Finding the best thresholds of FEV1 and dyspnea to predict 5-year survival in COPD patients: the COCOMICS study. *PLoS One*, 9:e89866.

Barrio, I., Arostegui, I., Rodríguez-Álvarez, M. X., and Quintana, J. M. (2016). A new approach to categorising continuous variables in prediction models: Proposal and validation. *Statistical Methods in Medical Research*, in press.

Bestall, J. C., Paul, E. A., Garrod, R., Garnham, R., Jones, P. W., and Wedzicha, J. A. (1999). Usefulness of the medical research council (MRC) dyspnoea scale as a measure of disability in patients with chronic obstructive pulmonary disease. *Thorax*, 54, 581–586.

Buist, A. S., Vollmer, W. M., and McBurnie, M. A. (2008). Worldwide burden of COPD in high-and low-income countries. Part I. The Burden of Obstructive Lung Disease (BOLD) Initiative. *The International Journal of Tuberculosis and Lung Disease*, 12, 703–708.

Celli, B. R., Cote, C. G., Marin, J. M., Casanova, C., Montes de Oca, M., Mendez, R. A., Pinto Plata, V., and Cabral, H. J. (2004). The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *New England Journal of Medicine*, 350, 1005–1012.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. CRC Press.

Esteban, C., Arostegui, I., Aburto, M., Moraza, J., Quintana, J. M., Aizpiri, S., Basualdo, L. V., and Capelastegui, A. (2014). Influence of changes in physical activity on frequency of hospitalization in chronic obstructive pulmonary disease. *Respirology*, 19, 330–338.

Esteban, C., Quintana, J. M., Aburto, M., Moraza, J., and Capelastegui, A. (2006). A simple score for assessing stable chronic obstructive pulmonary disease. *QJM - An International Journal of Medicine*, 99, 751–759.

Faraggi, D. and Simon, R. (1996). A simulation study of cross-validation for selecting an optmimal cutpoint in univariate survival analysis. *Statistics in Medicine*, 15, 2203–2213.

Fletcher, C. M., Elmes, P. C., Fairbairn, A. S., and Wood, C. H. (1959). The significance of respiratory symptoms and the diagnosis of chronic bronchitis in a working population. *British Medical Journal*, 2, 257.

Gönen, M. and Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92, 965–970.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81, 515–526.

Harrell, F. E. (2015). *rms: Regression Modeling Strategies*. R package version 4.3-0.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA: The Journal of the American Medical Association*, 247, 2543–2546.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61, 92–105.

Jones, R. C., Donaldson, G. C., Chavannes, N. H., Kida, K., Dickson-Spillmann, M., Harding, S., Wedzicha, J. A., Price, D., and Hyland, M. E. (2009). Derivation and validation of a composite index of severity in chronic obstructive pulmonary disease: the DOSE index. *American Journal of Respiratory and Critical Care Medicine*, 180, 1189–1195.

Lausen, B. and Schumacher, M. (1996). Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Computational Statistics & Data Analysis*, 21, 307–326.

Liu, X. and Jin, Z. (2015). Optimal survival time-related cut-point with censored data. *Statistics in Medicine*, 34, 515–524.

Marin, J. M., Alfageme, I., Almagro, P., Casanova, C., Esteban, C., Soler-Cataluña, J. J., de Torres, J. P., Martínez-Camblor, P., Miravitlles, M., Celli, B. R., and Soriano, J. B. (2013). Multicomponent indices to predict survival in COPD: the COCOMICS study. *European Respiratory Journal*, 42, 323–332.

Mebane, W. R. and Sekhon, J. S. (2011). Genetic optimization using derivatives: the rgenoud package for R. *Journal of Statistical Software*, 42, 1–26.

Meira-Machado, L. and Faria, S. (2014). A simulation study comparing modeling approaches in an illness-death multi-state model. *Communications in Statistics-Simulation and Computation*, 43(5), 929–946.

Mo, Q., Gonen, M., and Heller, G. (2012). *CPE: Concordance Probability Estimates in Survival Analysis*. R package version 1.4.4.

Pencina, M. J. and D'Agostino, R. B. (2004). Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in medicine*, 23(13), 2109–2123.

Pepe, M. S., Zheng, Y., Jin, Y., Huang, Y., Parikh, C. R., and Levy, W. C. (2008). Evaluating the roc performance of markers for future events. *Lifetime data analysis*, 14(1), 86–113.

Puhan, M. A., Garcia-Aymerich, J., Frey, M., ter Riet, G., Antó, J. M., Agustí, A. G., Gómez, F. P., Rodríguez-Roisín, R., Moons, K. G., Kessels, A. G., and Held, U. (2009). Expansion of the prognostic assessment of patients with chronic obstructive pulmonary disease: the updated bode index and the ado index. *The Lancet*, 374, 704–711.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rabe, K. F., Hurd, S., Anzueto, A., Barnes, P. J., Buist, S. A., Calverley, P., Fukuchi, Y., Jenkins, C., Rodriguez-Roisin, R., van Weel, C., and Zielinski, J. (2007). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: Gold executive summary. *American Journal of Respiratory and Critical Care Medicine*, 176, 532–555.

Rota, M., Antolini, L., and Valsecchi, M. G. (2015). Optimal cut-point definition in biomarkers: the case of censored failure time outcome. *BMC Medical Research Methodology*, 15, 24.

Schmid, M. and Potapov, S. (2012). A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Statistics in Medicine*, 31, 2588–2609.

Sima, C. S. and Gönen, M. (2013). Optimal cutpoint estimation with censored data. *Journal of Statistical Theory and Practice*, 7, 345–359.

Steyerberg, E. W., Moons, K. G. M., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., Altman, D. G., and Group, P. (2013). Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Medicine*, 10, e1001381.

Turner, E., Dobson, J., and Pocock, J. (2010). Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. *Epidemiologic Perspectives & Innovations*, 7, 9.

Vestbo, J., Hurd, S. S., Agustí, A. G., Jones, P. W., Vogelmeier, C., Anzueto, A., Barnes, P. J., Fabbri, L. M., Martinez, F. J., Nishimura, M., Stockley, R. A., Sin, D. D., and Rodriguez-Roisin, R. (2013). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: Gold executive summary. *American Journal of Respiratory and Critical Care Medicine*, 187, 347–365.

SORT 41 (1) January-June 2017, 93-116 DOI: 10.2436/20.8080.02.52

Bayesian correlated models for assessing the prevalence of viruses in organic and non-organic agroecosystems

Elena Lázaro[1], Carmen Armero[1] and Luis Rubio[2]

Abstract

Virus diseases constitute one of the most important limiting factors in horticultural production. Cultivation of horticultural species under organic management has increased in importance in recent years. However, the sustainability of this new production method needs to be supported by scientific research, especially in the field of virology. We studied the prevalence of three important virus diseases in agroecosystems with regard to its management system: organic *versus* non-organic, with and without greenhouse. Prevalence was assessed by means of a Bayesian correlated binary model which connects the risk of infection of each virus within the same plot and was defined in terms of a logit generalized linear mixed model (GLMM). Model robustness was checked through a sensitivity analysis based on different hyperprior scenarios. Inferential results were examined in terms of changes in the marginal posterior distributions, both for fixed and for random effects, through the Hellinger distance and a derived measure of sensitivity. Statistical results suggested that organic systems show lower or similar prevalence than non-organic ones in both single and multiple infections as well as the relevance of the prior specification of the random effects in the inferential process.

MSC: 62-07; 62F15; 62J12; 62P10; 62P12.

Keywords: Hellinger distance, model robustness, risk infection, sensitivity analysis, virus epidemiology.

1. Introduction

Society is becoming increasingly concerned about environmental damage caused by agricultural activities. The sustainability of conventional agriculture is now being ques-

[1] Department of Statistics and Operations Research, Faculty of Mathematics, Universitat de València, Spain. elena.lazaro@uv.es, carmen.armero@uv.es

[2] Valencian Institute for Agricultural Research, Spain. lrubio@ivia.es

Received: June 2016
Accepted: December 2016

tioned, which is prompting traditional production systems to evolve toward production methods that can protect both environmental and human health (Van Bruggen, 1995; Bengtsson et al., 2005).

In recent decades, organic agriculture has grown rapidly in comparison with other agricultural systems. The adoption of these new agricultural practices has brought about the need to compare low-input and conventional systems to verify whether agroecosystem sustainability can be achieved (Bettiol et al., 2004). Despite the emergence of organic agriculture systems, the literature on their effects and interactions is scarce and insufficient, above all in the field of virology (Tomlinson, 1987). Diseases caused by viruses constitute a major threat to the large-scale production of crops worldwide, causing serious economic losses and undermining sustainability (Gallitelli, 2000). Assessing the risk of infection should therefore be a priority in the study of the epidemiology of such virus diseases.

The ecological and epidemiological factors that determine virus infections in vegetable crops are diverse and little is known about them. The sources and spread of viruses, together with certain agricultural and horticultural practices, have a strong influence on their prevalence (Hanssen et al., 2010). In this respect, studies on the risk of virus infections need to characterize the agroecosystem balance as well as understand the complex relationships between organisms (plants, pathogens, and vectors) and environment (Serra et al., 1999).

The main scientific question addressed in this paper is the study and comparison of the risk of different virus infections in tomato and pepper plots characterized by their agroecosystem. Specifically, we focus on the detection and quantification of the effects associated with organic management. The agroecosystem of each plot is defined through a set of covariates containing information on its management conditions and altitude. Agroecosystems are dynamic entities (Finley et al., 2011) with complex sources of uncertainty and hierarchies. Following Thornley and France (2007), the estimation of the infection risk of different viruses within the same plot would require the modelling of not only a suitable set of covariates but also the inclusion of some probabilistic terms which connect the different observations of the same individidual.

The inclusion of dependence and/or correlation relationships among variables, response and/or covariates, is usually done by means of random effects whose stochastic nature adds much more probability to the structure of the model. Bayesian reasoning provides a natural environment for analysing them mainly because of the own conception of the Bayesian probability theory, which specifies all the uncertainties in the model through probabilistic elements (Loredo, 1990). Some applied papers that illustrate the benefits of hierarchical Bayesian models in biometrics scenarios are Alvares et al. (2016) in agriculture, Paradinas et al. (2015) in fisheries, Paciorek et al. (2009) in forestry, and Clark et al. (2007) in ecology.

A Bayesian binary correlated model under the generalized linear mixed models (GLMM) specification was considered to perform a regression analysis of the prevalence of the different viruses. Random effects were used to correlate the risk of infection of each virus in the same plot and quantify the intra-plot ability to be infected. Robustness in hierarchical Bayesian models is a major concern as it can be affected by an inappropriate choice of the hyperprior distributions for hyperparameters (Lambert et al., 2005; Gelman, 2006; Roos and Held, 2011; Roos et al., 2015). To this effect, the sensitivity of the modelling was tested using several specifications for the hyperprior distribution of the random effects scale parameter. A general measure based on the Hellinger distance (Le Cam, 2012), with its calibration, was used to quantify discrepancies in the subsequent posterior marginal distribution of the common regression coefficients and hyperparameter.

The remainder of this article is organized as follows: Section 2 reviews the data and presents the formulation of the model. Section 3 reports and discusses the results with regard to multiple and single viral infections. Section 4 proposes several random effects specifications and analyses the robustness of the estimated models through a sensitivity measure based on the Hellinger distance. Some concluding remarks are given in Section 5.

2. Viruses data and statistical modelling

2.1. Data description

Globally, about 30 viruses are capable of affecting the most known horticultural crops. However, despite being able to infect a wide variety of species, they usually affect Solanaceae species, specially tomato (*Solanum lycopersicum*) and pepper (*Capsicum annuum L.*). These species are two of the most common vegetable crops grown in Spain whose production is being seriously limited by virus diseases. There has recently been a considerable increase in the cultivation of these vegetables under integrated systems such as organic agriculture. It is therefore essential to carry out subsequent virus prevalence studies in order to guarantee their sustainability.

A project under the auspices of the Valencian Institute Agricultural Research was conducted in the summer of 2012 in the Valencian region for this purpose. A total of 30 plots in tomato and pepper production were selected according to their system of production. Each plot was evaluated in terms of its agroecosystem characterization and the presence or absence of three different viral infections in the crops: tomato mosaic virus (ToMV), cucumber mosaic virus (CMV) and tomato spotted wilt virus (TSWV). These viruses affect both tomato and pepper crops equally, are transmitted in different ways, and can cause substantial economic losses. The presence of each specific virus infection in a plot was assumed when the virus was detected in at least one of eight randomly-selected plants. The enzyme-linked immunosorbent assay (ELISA) technique (Clark et al., 1976) was used to detect each virus.

The assessment of the agroecosystem of each plot was determined by its management condition and altitude. Management condition was evaluated by classifying each plot as organic, non-organic with greenhouse structure, and non-organic with no greenhouse structure. These categories were defined according to the most representative agroecosystems in Spanish agriculture. Organic plots differ from the non-organic ones in many respects, but substantial differences are related to the use of agrochemicals and other external inputs with important influence in pest and disease prevalence. In fact, some purported drawbacks related to organic agriculture include an increasing incidence of pest damage and higher risks of pest outbreaks (Letorneau and Goldstein, 2001). All plots classified as organic complied with the current regulation and were certificated as such by the Organic Agriculture Committee of the Autonomous Government of València. The presence of greenhouse in non organic plots was also considered because is a frequent practice in non-organic systems. The use of covering protections suppose a physical barrier which is directly related to virus infection in the sense that denies insects (vector of virus transmission) acces to plants.

Of the total of 30 plots of our study, 18 were classified as organic and 12 as non-organic, 5 of them with greenhouse structure. For organic plots, the proportion of infected plants with ToMV, CMV, and TSWV was 0.222, 0.167, and 0.056, respectively. In the case of non-organic plots with greenhouse these proportions were 0.400, 0.200, and 0.200, respectively, and 0.143, 0.286, and 0.286 for non-organic plots without greenhouse. The organic plots presented a lower proportion of plants infected by CMV and TSWV viruses, but the prevalence of ToMV was lowest in the non-organic plots with no greenhouse.

2.2. Statistical model

We consider a logit GLMM for correlated binary responses (Ntzoufras, 2009) to model the Bernoulli random variable Y_{ij} which describes the presence or absence of virus j ($j = 1$ corresponds to ToMV, $j = 2$ to CMV, and $j = 3$ to TSWV) in plot i,

$$\begin{aligned} (Y_{ij} \mid \theta_{ij}) &\sim \text{Bernoulli}(\theta_{ij}), \\ \text{logit}(\theta_{ij}) &= x_i^{\mathrm{T}}\boldsymbol{\beta}_j + b_i,\ i = 1, \dots, 30, \end{aligned} \tag{1}$$

where θ_{ij} is the probability that virus j will be detected in plot i and represents risk of infection; $\boldsymbol{x}_i$ is the vector of covariates; $\boldsymbol{\beta}_j$ is the corresponding vector of the regression coefficients; and $(b_i \mid \sigma_b^2) \sim \text{N}(0, \sigma_b)$ is a normal random effect associated with plot i with mean zero and standard deviation σ_b. The three management conditions were coded in a sequence of two dummy variables (organic and non-organic, with and without greenhouse structure) to avoid overparameterization, with organic management as the reference category.

Random effects capture within-plot variability and correlate prevalence among all viruses so that each individual virus infection is determined by its own agroecosystem effect and an individual effect plot which denotes its ability to be infected. They also provided conditional independence among the prevalence of the three viruses as follows

$$P(Y_{ij} = y_j, j = 1,2,3 \mid \boldsymbol{\beta}, b_i, x_i) = \prod_{j=1}^{3} P(Y_{ij} = y_j \mid \boldsymbol{\beta}_j, b_i, x_i), \tag{2}$$

where $y_j \in \{0,1\}$, $j = 1,2,3$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_2)^{\mathsf{T}}$, and the conditional probability that plot i will be infected with virus j can be expressed as

$$P(Y_{ij} = 1 \mid \boldsymbol{\beta}_j, b_i, x_i) = \frac{\exp\{x_i^{\mathsf{T}}\boldsymbol{\beta}_j + b_i\}}{1 + \exp\{x_i^{\mathsf{T}}\boldsymbol{\beta}_j + b_i\}}, \tag{3}$$

The joint marginal distribution obtained integrating out the random effects in (4),

$$P(Y_{ij} = y_j, j = 1,2,3 \mid \boldsymbol{\beta}, \sigma_b, x_i) = \int P(Y_{ij} = y_j, j = 1,2,3 \mid \boldsymbol{\beta}, b_i, x_i)\, \mathrm{N}(b_i \mid 0, \sigma_b)\, \mathrm{d}b_i, \tag{4}$$

does not depend on the subject-specific random effects and can be interpreted as the common risk infection of a generic plot from the population with the same agroecosystem and altitude.

Inference was carried out using Bayesian statistics. We therefore needed to elicit a prior distribution for the parameters and hyperparameters to complete the Bayesian model. We considered a prior independent default scenario with normal distributions centered at zero and a wide variance for the regression coefficients. As previously introduced, the specification of a hyperprior distribution for the random effects scale parameter is a challenging issue (Lambert et al., 2005; Gelman, 2006; Roos and Held, 2011; Roos et al., 2015). Section 4 contains a sensitivity analysis of the performance of various traditional hyperprior choices (gamma, uniform and half-normal) in our study. This analysis led us to choose the uniform distribution $\mathrm{Un}(\sigma_b \mid 0, 100)$ for the standard deviation of the random effects. Consequently

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma_b) &= \textstyle\prod_{j=1}^{3} \prod_{k=0}^{3} \pi(\beta_{jk})\, \pi(\sigma_b) \\ &= \textstyle\prod_{j=1}^{3} \prod_{k=0}^{3} \mathrm{N}(\beta_{jk} \mid 0, \sigma^2 = 1000)\, \mathrm{Un}(\sigma_b \mid 0, 100) \end{aligned} \tag{5}$$

where $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \beta_{j2}, \beta_{j3})^{\mathsf{T}}$ are the regression coefficients associated with organic, non-organic with and without greenhouse and altitude (in logarithmic scale) for virus j.

3. Results

The posterior distribution $\pi(\boldsymbol{\beta}, \sigma_b \mid \mathscr{D})$, where $\mathscr{D}$ denotes data, was approximated using Markov chain Monte Carlo (MCMC) simulation methods with WinBUGS Software (Lunn et al., 2000). Random effects models, and Bayesian categorical GLMs in particular, involve many computational difficulties (Albert and Chib, 1993). We fixed the number of iterations and the burn-in period with very large values to avoid strong correlation in the MCMCs samples and get a reliable sample of the posterior distribution. Specifically, simulation was run considering three Markov chains with 1 000 000 iterations and a burn-in period with 100 000. In addition, the chains were thinned by storing every 10th iteration in order to reduce autocorrelation in the saved sample and avoid computer memory problems.

Trace plots of the simulated values of the chains appear overlapping one another, indicating stabilization. Convergence of the chains to the posterior distribution was assessed using the potential scale reduction factor, $\hat{R}$, and the effective number of independent simulation draws, neff. In all cases, the $\hat{R}$ values were equal or close to 1 and neff > 100, thus indicating that the distribution of the simulated values between and within the three chains was practically identical, and that sufficient MCMC samples had been obtained, respectively (Gelman and Rubin, 1992).

3.1. Management conditions

Multiple viral infections that may result in synergisms or antagonisms are frequently found in nature, with unpredictable pathological consequences. Synergistic interactions resulting from mixed infections with two or more viruses are common and well documented in plants (García-Cano et al., 2006). Viral synergism could affect various growth variables such as plant height, weight, and yield (Murphy and Bowen, 2006), and in extreme cases can lead to plant death.

The joint posterior distribution, $\pi(P(Y_{ij} = y_j, j = 1,2,3 \mid \boldsymbol{\beta}, \sigma_b, x_i) \mid \mathscr{D})$, where $y_j \in \{0,1\}$, of the risk infection given in (4) for a generic plot at given altitude in each of the management systems is the basic tool for assessing such synergisms and antagonisms. This posterior distribution is also the starting point for the computation of relevant conditional or marginal inferences.

We begin by discussing some results about multiple viral infections with regard to plot management condition: the posterior distribution of the prevalence of the total number of viruses in a plot and the posterior distribution of the risk of a third infection in plots already infected with two of the viruses. Figure 1a shows the mean of the posterior distribution associated to the presence of 0, 1, 2 and 3 viruses in a generic plot i located at 76 meters of altitude (the sample mean) with regard to its management system. Most of the plots have no infections, but the organic ones present the highest rates for plots without infections. Non-organic plots, with and without greenhouse, behave similarly.

Figure 1b shows the posterior mean of the risk of a third infection in plots already infected with two of the viruses. Outcomes are also obtained for a generic plot i situated at 76 meters of altitude (the sample mean) with regard to its management system. For condition ToMV in the presence of CMV and TSWV, organic and non-organic with greenhouse plots behave similarly with probabilities around 0.6. This is not the case for non-organic with no greenhouse plots, with an estimated probability close to 0.2. CMV infection given ToMV and TSWV presents homogeneous results in all manage-

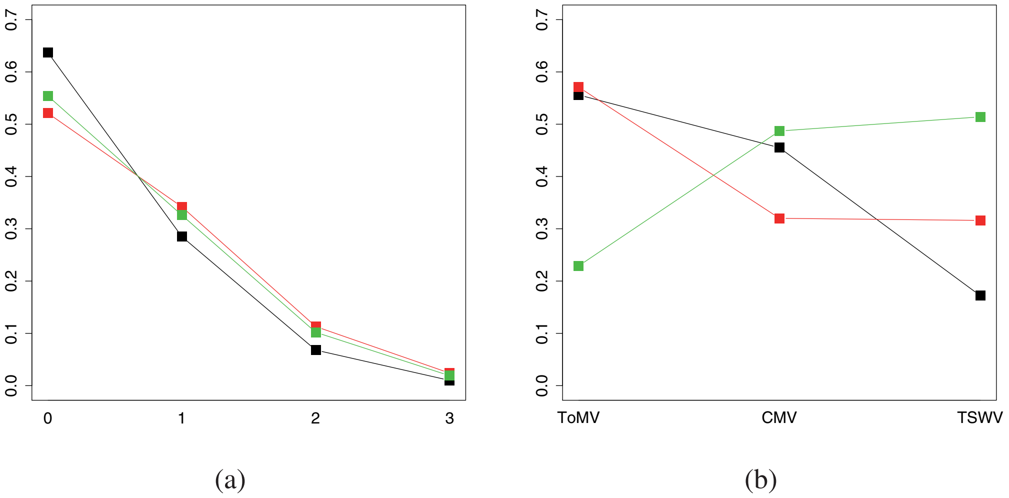


Figure 1: *(a) Probability (mean of the posterior distribution) for the presence of 0, 1, 2 and 3 viruses in organic (black), non organic-green (red) and non organic-non green (green) management systems. (b) Probability (mean of the posterior distribution) of the risk of a third infection in plots already infected with two of the viruses in organic (black), non organic-green (red) and non organic-non green (green) management systems.*

ment systems, with a higher difference among estimated probabilities of 0.167. The pattern for the probability of a TSWV infection in plots already infected with ToMV and CMV seems to be different among the management conditions: non-organic with no greenhouse systems shows the highest probability (0.514), followed by non-organic with greenhouse plots (0.316), and organic (0.172), respectively. It is difficult to detect a general trend on conditional infections among the different agroecosystems analysed. This is a very interesting subject and surely a new study with more data would be necessary in order to better understand them.

The marginal effect of the management conditions in each virus was assessed through the marginal posterior distribution $\pi(P(Y_{ij} = 1 \mid \boldsymbol{\beta}, \sigma_b, x_i) \mid \mathscr{D})$. Table 1 shows a descriptive of the posterior distribution of the risk of infection for each virus and management conditions for a generic plot situated at a height of 76 meters (the sample median). The lowest risk of infection for a generic plot under organic management is for TSWV virus. The most relevant differences among the management conditions were found for virus ToMV. In contrast, virus CMV seemed the most stable. However, the organic effect was weaker for ToMV risk, approximately about four times the one for TSWV virus. It is important to mention the great uncertainty associated to all marginal posterior distributions in the analysis, mainly due to the combination of the reduced size of the sample and the usual scarce information of binary data. To this effect, a bigger experiment would be necessary for a more informative and objective study that allows to reach more precise conclusions about the subject.

Table 1: *Summary of the posterior distribution of the risk of infection for each management condition and virus.*

Virus	Management	Mean	Sd	$Q_{2.5\%}$	$Q_{50\%}$	$Q_{97.5\%}$
ToMV	Organic	0.225	0.184	0.008	0.181	0.734
	Non-organic, greenhouse	0.311	0.252	0.006	0.248	0.900
	Non-organic, no greenhouse	0.100	0.147	0.000	0.041	0.553
CMV	Organic	0.169	0.161	0.004	0.124	0.634
	Non-organic, greenhouse	0.155	0.190	0.001	0.080	0.719
	Non-organic, no greenhouse	0.234	0.216	0.004	0.168	0.809
TSWV	Organic	0.057	0.093	0.000	0.026	0.309
	Non-organic, greenhouse	0.174	0.203	0.001	0.095	0.764
	Non organic, no greenhouse	0.253	0.223	0.005	0.189	0.831

Comparison of the three management systems was also quantified with the posterior distribution of the risk difference (*RD*) (Christensen et al., 2011). *RD* is an absolute and intuitive measure of association for quantifying difference between proportions associated to an outcome of interest in two groups. It is defined in $[-1,1]$ so that $RD = 0$ means no difference between groups, $-1 \leq RD < 0$ that risk is greater in group 2, and $0 < RD \leq 1$ the opposite.

Figure 2 shows, for each virus, the posterior mean and 95% credible interval of the *RD* between organic and non-organic, with and without greenhouse, generic plots. Information provided by this graphic reaffirms the results in Table 1. Note that the differences between organic management conditions and the two non-organic conditions are clear in the case of TSWV infection: both posterior distributions are highly concentrated on the negative *RD* values with associated posterior probabilities 0.764 and 0.910 when com-

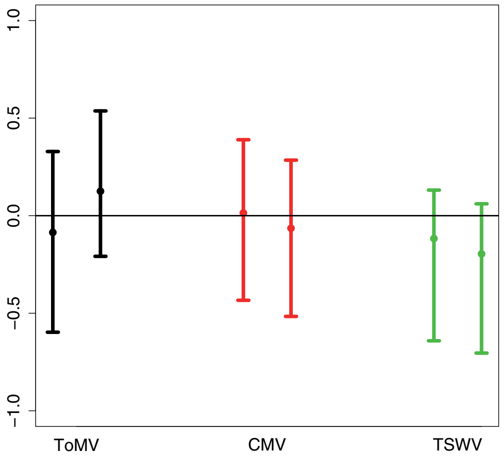


Figure 2: *Posterior mean and 95% credible interval of the RD between organic system in relation to non organic-green (left) and non organic-no green (right) system for ToMV, CMV and TSWV infections.*

paring organic and non-organic with and without greenhouse management, respectively. For CMV infections, the results are less clear, with posterior probabilities of 0.395 and 0.611, respectively. In the case of ToMV infection, there are few differences between organic and non-organic with greenhouse conditions (posterior probability of a negative difference is 0.620), but a relevant probability, 0.84, that the risk of infection will be greater in organic than in non-organic without greenhouse.

3.2. Altitude condition effect

Plot altitude is a relevant epidemiological information due to its important role in shaping insect vector distributions and virus survival. The effect of altitude on the risk of infection is clearly negative in all viruses and therefore we can expect a decrease of the risk of infection as altitude increases. Figure 3a shows the posterior distribution of the regression coefficient associated to altitude for each virus: -0.914, -0.745 and -0.480 are, respectively, the subsequent posterior mean of the coefficient for virus ToMV, CMV, and TSWV, with posterior probabilities 0.940, 0.904, and 0.768 associated to their negative values. Note that virus ToMV is the most negatively associated with altitude. Figure 3b shows the posterior distribution of the *RD* between two generic organic plots with altitudes of 16 and 604 m, the lowest and highest values of the organic plots in the sample. These graphics are in line with the previous comments and also indicate the less important role of altitude in the risk of a TSWV infection in organic crops.

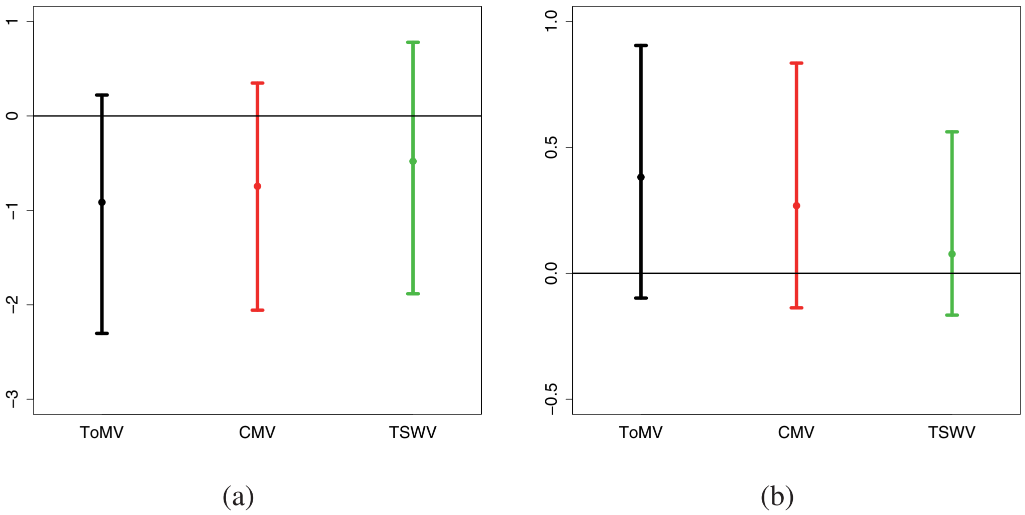


Figure 3: *For virus ToMV (in black), CMV (in red), and TSWV (in green): posterior mean and 95% credible interval of the regression coefficient associated to the altitude (in logarithmic scale) (a), and posterior distribution of the RD between a typical organic plot at altitudes 16 and 604 m (b).*

3.3. Individual random effects

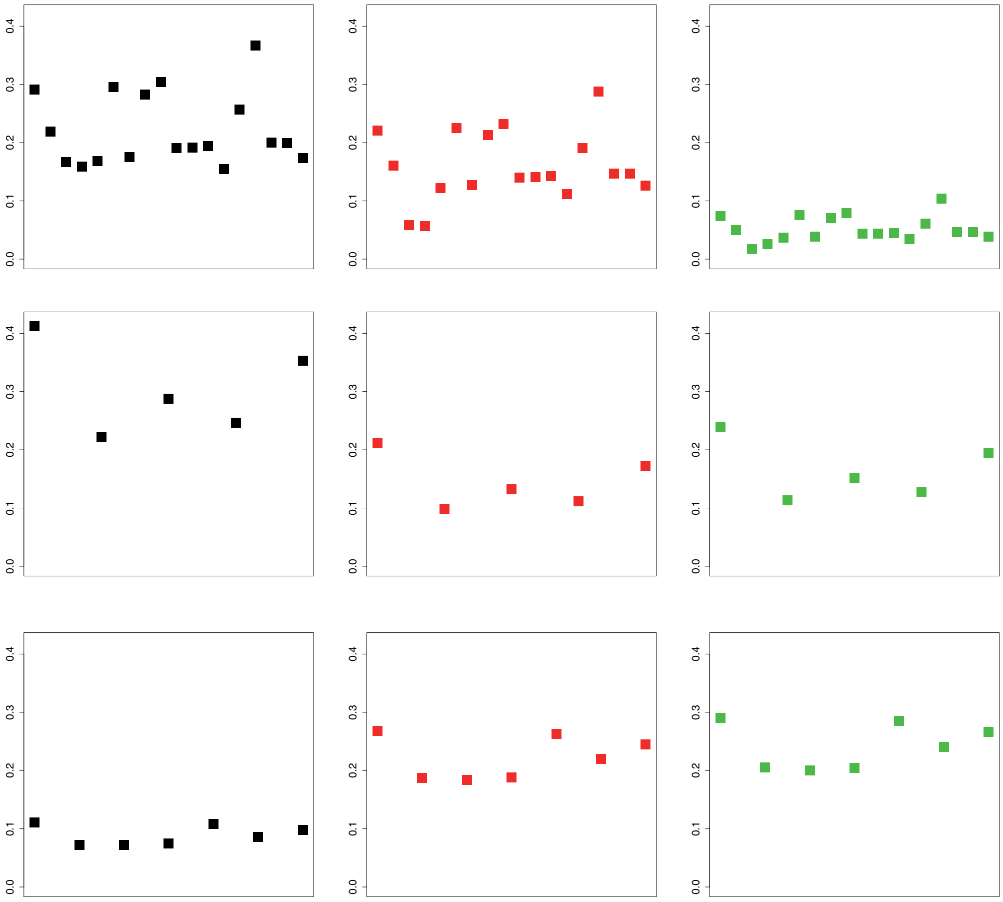
Random effects for each plot capture the ability to be infected of individual plots, thus correlating the risk of infection among the viruses of each plot. Since each individual random effect is responsible for the differences in the estimation of the risk between plots managed under similar agroecosystem conditions, quantifying their contribution to the analysis in terms of factors and covariates is highly relevant to our understanding of the weight of the common and individual elements in the model.

The mean of the posterior distribution of the standard deviation, σ_b, of the plot random effect is 0.968 with a 95% credible interval [0.046, 2.671]. In addition, we assessed the contribution of the random effect associated to each plot towards the conditional posterior distribution of the risk of infection $\pi(P(Y_{ij} = 1 \mid \boldsymbol{\beta}, b_i, x_i) \mid \mathscr{D})$. It was estimated individually for the three viruses at the altitude of 76 meters with the purpose of assessing differences in risk infection among individuals that share the specification of the vector of covariates x_i, that is to say, plots that were managed under the same system. Figure 4 shows a mosaic of subfigures in which each one displays the posterior expectation of the risk of infection for each plot grouped according to management condition (rows) and the type of virus infection (columns).

We can distinguish a certain stability in risk infection regarding individuals belonging to non-organic no greenhouse systems (row 3) with maximum differences among individuals of 0.039, 0.084 and 0.090 for ToMV, CMV and TSWV respectively. Non-organic with greenhouse plots (row 2) are less similar with maximum differences in risk infection no greater than 0.190 (ToMV). Organic plots showed the most remarkable differences among their individuals, with maximum differences of 0.211 for ToMV and 0.231 for CMV. In contrast TSWV showed the opposite behaviour with a slight maximum difference of 0.087. In conclusion, we suspect the strong relevance of the common elements in the model (fixed effects) in the case of non-organic and no greenhouse plots regardless of virus infection. On the other hand, in the case of organic plots the weight of the common elements effect in the model was not so evident considering that not all viruses exhibited a similar tendency: ToMV and CMV risk infection varied considerably among individuals, but this was not the case with TSWV.

4. Sensitivity analysis

Bayesian GLMMs are a particular class of models for which the estimation process can be seriously affected by the elicitation of prior distributions for the random effects scale parameter (standard deviation, σ_b, or a one-to-one transformation of it, variance σ_b^2 or precision $\tau_b = 1/\sigma_b^2$). Special attention is required in studies where the number of groups is small, σ_b is close to zero, and/or the number of groups is large compared to the number of observations in each group (Box and Tiao, 1992; Gelman, 2006; Roos and Held, 2011). This latter situation is the case of our study, with $I = 30$ plots and only three observations in each of them. An additional element that aggravates the situation is the sparsity of the data due to its categorical, binary condition. We conducted a sensitivity analysis of the posterior distribution to the specification of several prior hyperdistributions for the random effects scale parameter. This analysis was based on the methodology developed in McCulloch (1989), Roos and Held (2011), and Roos et al. (2015) regarding the stability of the marginal posterior distribution of the regression coefficients of the model and the relative changes in the subsequent marginal posterior distributions of the random effects scale parameter.



(a) ToMV (b) CMV (c) TSWV

Figure 4: *Posterior mean of the conditional posterior distributions associated to management systems organic (row 1), non organic and greenhouse (row 2) and non organic and non greenhouse (row 3) for viruses ToMV (column1), CMV (column 2) and TSWV (column 3) obtained from a fixed altitude value of 76 m.*

4.1. Hyperprior distributions

For the random effects scale parameter, different hyperprior distributions were specified for τ_b within the family of gamma, and for σ_b within uniform and half-normal distributions

- Gamma: $\text{Ga}(0.001, 0.001)$, $\text{Ga}(0.005, 0.005)$, and $\text{Ga}(0.05, 0.05)$ (Ga1, Ga2, and Ga3, respectively),
- Uniform: $\text{Un}(0, 100)$, $\text{Un}(0, 55.63)$, and $\text{Un}(0, 7.92)$ (Un1, Un2, and Un3), and
- Half-normal: $\text{HN}(10)$, $\text{HN}(3.0387)$, and $\text{HN}(0.3965)$ (HN1, HN2, and HN3).

Gamma distributions were parameterized in terms of a shape and a rate parameter, and half-normal distributions were set according its standard deviation. Hyperdistributions Ga1, Un1, and HN1 were considered the default choices due to their "noninformative" nature and their common use in Bayesian applications. In addition, two other hyperparameter specifications within each family of hyperdistributions were contemplated to assess the effect of small and medium perturbations in the hyperparameter specifications on posterior inferences. These hyperprior distributions were set following the criterion of the Hellinger distance (Le Cam, 2012). This is a symmetric and invariant measure of discrepancy between two probability distributions taking values between 0 and 1, where the value 0 represents no divergence and 1, full divergence (See Appendix 1).

Hyperparameter values were assessed considering two reference Hellinger distance values, a small and a medium perturbation. This computation was based on the analytical expression of the Hellinger distance between gamma, uniform and half-normal distributions (see Appendix 1). Small perturbation was associated to a Hellinger distance of 0.541 and medium to 0.848. Consequently, Ga2, Un2, and HN2 hyperparameteres were determined to obtain a Hellinger distance of 0.541 in relation to hyperdistributions Ga1, Un1, and HN1, respectively. Hyperparameter values for Ga3, Un3, and HN3 were selected because of their Hellinger distance, 0.848, to hyperpriors Ga1, Un1, and HN1, respectively.

Focusing on gamma hyperdistributions, Ga1 exhibits the widest range of uncertainty with a variance of 1000. It is frequently used in many of the examples provided with the WinBUGS software (Lunn et al., 2012) and shows a uniform shape for most of the range with a spike of probability density near zero. Ga2 and Ga3 share this shape, although they show lower range coverage as a consequence of their fewer variance values, 200 and 20. Hyperprior Un1 is recommended by Spiegelhalter et al. (2004) in their book on clinical trials. It is a very generous distribution allowing for a great space of values due to its variance of 833.3. Un2 and Un3 display variance values of 257.84 and 5.23, and as such they are very different from the non-null density range. The half-normal default option, HN1, is a choice used in Thompson et al. (1997) and Roos and Held (2011). It exhibits a variance of 36.3 giving a low probability to values greater than this. HN2 and HN3 are more informative versions, especially HN3 with a variance value of 0.06.

We conducted nine independent inferential processes with the same data and the same marginal prior distribution $\pi(\boldsymbol{\beta})$ for the regression coefficients as in (5) but varying marginal hyperprior distribution according to the specifications previously presented.

4.2. Sensitivity of the regression coefficients

We discuss sensitivity of the marginal posterior distributions of the regression coefficients derived from the inferential processes described above. Discrepancies among the estimates of posterior marginal distributions were the result of alterations in the hyperprior values. Hellinger distances between posterior marginal distributions approximated by MCMC methods were computed via expression (A.1) in Appendix 1 and implemented by means of the function `HDistNoSize` in the `R` package `bmk` (Krachey and Boone, 2012). Furthermore, to facilitate interpretation these values were calibrated with regard to a normal distribution with variance 1 (see Appendix 2 for more details about calibration).

Table 2 shows the calibration of the Hellinger distance between the posterior marginal distribution of the different coefficients of regression computed from the hyperpriors considered. In none of the comparisons the discrepancies observed were greater than the differences between the normal distributions $\mathrm{N}(0,1)$ and $\mathrm{N}(0.284,1)$, which reveals that Hellinger values are in general close to zero (see Table 4 in Appendix 2 where a calibration of the normal mean related to its subsequent Hellinger distance is displayed). Uniform distributions have the smallest discrepancies despite the existing differences among hyperpriors Un1, Un2, and Un3. The behaviour of half-normal distributions was similar to that of the uniform distributions in the case of hyperpriors HN1 and HN2. Nevertheless, inference from hyperprior HN3 exhibited the greatest discrepancies, surely due to its informative nature. Gamma showed greater discrepancies than uniform hyperpriors in all cases, although in none of the scenarios did these differences exceed those obtained from hyperprior HN3. Thus, the above comments enable us to conclude that our assumptions on the choice of hyperparameter prior distribution influences the estimates of the regression coefficients only to a minor extent.

We now discuss the effect of the different hyperpriors considered on the posterior distribution of each regression coefficient. Figure (5) is a mosaic of subfigures. Each subfigure displays the posterior mean of the regression coefficients of the different inferential processes conducted. The order of the points corresponds to the order in which hyperpriors are presented (Ga1, Ga2, Ga3; Un1, Un2, Un3; and HN1, HN2, HN3). A great similarity can repeatedly be seen, in practically all coefficients and viruses, between results from hyperpriors HN1 and HN2, and also those from the uniform hyperpriors. As expected, results from HN3 are very different, most likely due to its informative characteristics. Finally, posterior means from the analyses based on the gamma hyperpriors vary the most, indicating a greater sensitivity to parameter specification.

Table 2: *Calibration of the Hellinger distance between the posterior marginal distribution of the coefficients of regression associated to organic (β_o), non-organic with greenhouse ($\beta_{no\text{-}g}$), non-organic without greenhouse ($\beta_{no\text{-}ng}$) and altitude in logarithmic scale (β_{alt}) computed from hyperprior distributions Ga1 and Ga2, Ga1 and Ga3, Un1 and Un2, Un1 and Un3, HN1 and HN2, and HN1 and HN3.*

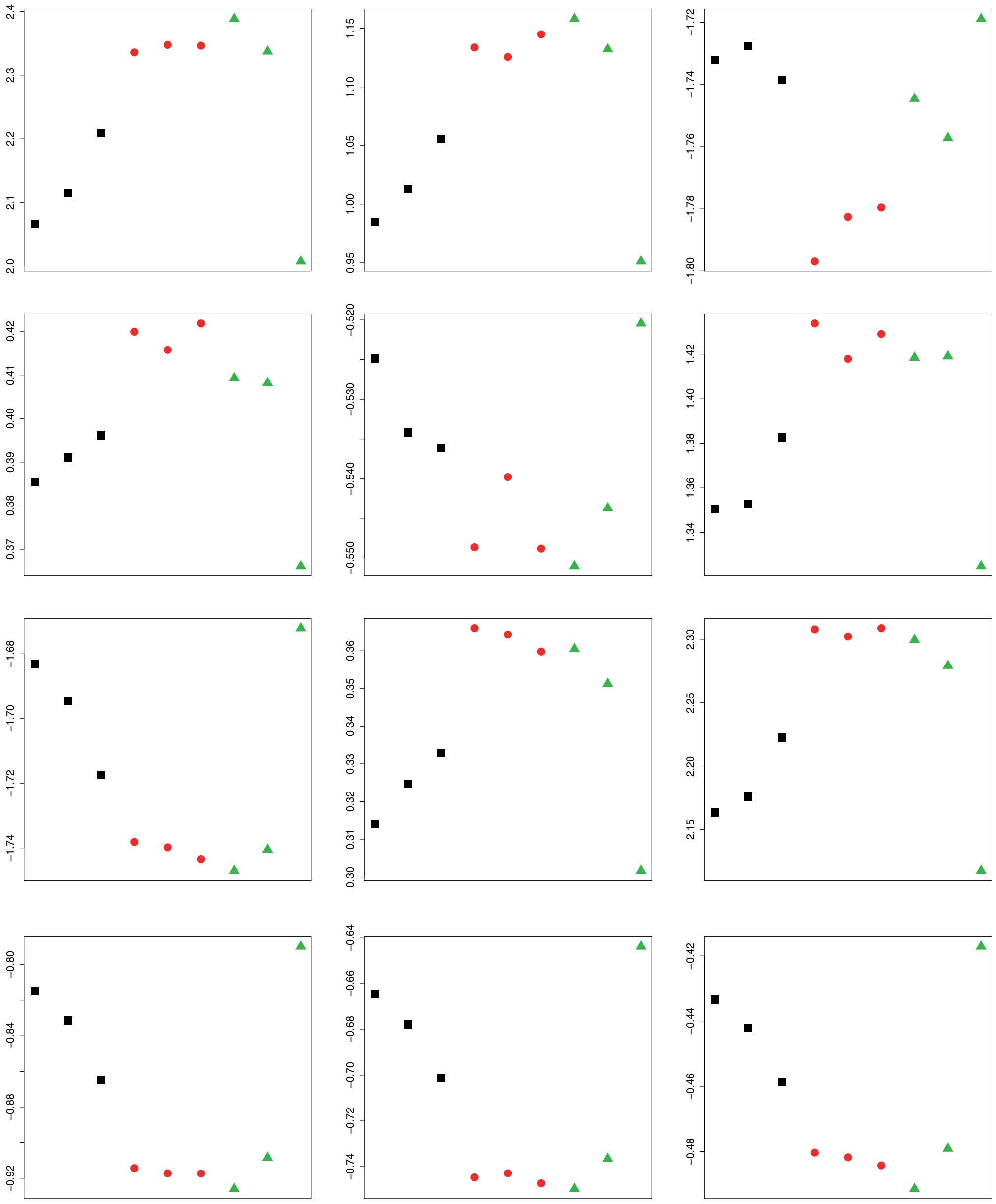
Virus	Coeff.	(Ga1,Ga2)	(Ga1,Ga3)	(Un1,Un2)	(Un1,Un3)	(HN1,HN2)	(HN1,HN3)
ToMV	β_{o}	0.038	0.084	0.024	0.022	0.034	0.236
	$\beta_{\text{no-g}}$	0.032	0.068	0.019	0.019	0.035	0.197
	$\beta_{\text{no-ng}}$	0.020	0.042	0.018	0.020	0.024	0.124
	β_{alt}	0.043	0.099	0.022	0.024	0.039	**0.284**
CMV	β_{o}	0.033	0.068	0.023	0.021	0.034	0.201
	$\beta_{\text{no-g}}$	0.029	0.056	0.021	0.019	0.025	0.148
	$\beta_{\text{no-ng}}$	0.029	0.060	0.019	0.020	0.027	0.171
	β_{alt}	0.037	0.085	0.023	0.023	0.038	0.249
TSWV	β_{o}	0.022	0.052	0.019	0.021	0.030	0.144
	$\beta_{\text{no-g}}$	0.024	0.043	0.021	0.020	0.025	0.108
	$\beta_{\text{no-ng}}$	0.023	0.048	0.020	0.019	0.025	0.139
	β_{alt}	0.028	0.069	0.020	0.019	0.034	0.193

4.3. Sensitivity of the variability of the random effects

We now discuss and assess the sensitivity of the random effects scale parameter corresponding to the inferential processes described in Subsection 4.1. Figure 6 shows the posterior marginal distribution (mean and 95% credible intervals) of the standard deviation of the random effects. It is worth noting that in the case of the gamma hyperpriors, the posterior marginal distribution $\pi(\sigma_b \mid \mathscr{D})$ is computed from the joint posterior $\pi(\boldsymbol{\beta}, \tau_b \mid \mathscr{D})$, which is based on the prior $\pi(\boldsymbol{\beta}, \tau_b)$. The results from the uniform hyperdistribution are stable, since the subsequent marginal posterior distributions are virtually indistinguishable. The opposite occurs for results from the gamma hyperpriors, with very different posterior distributions greatly influenced by the spike near zero of the subsequent hyperprior. The half-normal distribution also exhibits a sensitive performance, with the posterior distributions from HN1 and HN2 practically equal to those from the uniform distribution. As previously noted, the exception is for the posterior distribution from the informative HN3.

Finally, we used a sensitivity measure developed in Roos and Held (2011) to evaluate the relative change in the posterior marginal distribution of the random effects scale parameter with regard to subsequent change in the prior distribution. Changes in both prior and posterior distributions are assessed through the ratio between two Hellinger metrics in the form

$$S(\pi_1, \pi_2) = \frac{H(\pi_1(\theta \mid \mathscr{D}), \pi_2(\theta \mid \mathscr{D}))}{H(\pi_1(\theta), \pi_2(\theta))},$$



(a) ToMV (b) CMV (c) TSWV

Figure 5: *Posterior mean of the regression coefficients associated to plot categories organic (row 1), non organic and greenhouse (row 2), non organic and non greenhouse (row 3), and covariate altitude in logarithmic scale (row 4) for viruses ToMV (column 1), CMV (column 2), and TSWV (column 3) obtained from the full inferential process based on G1, G2 and G3 (black), Un1, Un2 and Un3 (red) and HN1, HN2 and HN3 (green) hyperpriors.*

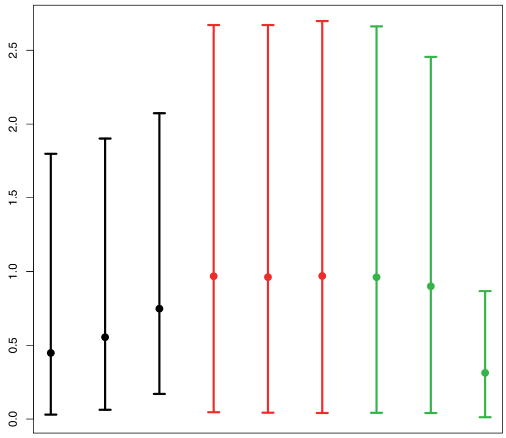


Figure 6: *Posterior mean and 95% credible interval for σ_b obtained from hyperpriors Ga1, Ga2, and Ga3 in black, Un1, Un2, and Un3 in red, and HN1, HN2, and HN3 in green.*

where $\pi_1(\theta \mid \mathscr{D})$ and $\pi_2(\theta \mid \mathscr{D})$ are the subsequent posterior distributions from $\pi_1(\theta)$ and $\pi_2(\theta)$. Note that $S(\pi_1, \pi_2)$ only depends on the Hellinger distance, and consequently, because of its invariancy to any one-to-one transformations we can parameterize the prior and posteriors in terms of τ_b or σ_b.

As expected, sensitivity values with gamma hyperpriors are very relevant,S(Ga1,Ga2) = 0.274 and S(Ga1,Ga3) = 0.477, with calibrated values 0.267 and 0.436 respectively. Thus, considering a Hellinger priors difference such as that reported between the normal distributions $\mathrm{N}(0,1)$ and $\mathrm{N}(1,1)$, their corresponding Hellinger posteriors difference should be understood as equal to that generated between the pair $\mathrm{N}(0,1)$ and $\mathrm{N}(0.267,1)$ in the case of hyperpriors Ga1 and Ga2, $\mathrm{N}(0,1)$ and $\mathrm{N}(0.436,1)$ in the case of Ga1 and Ga3 (see Appendix 2 for more details of calibration). In contrast, sensitivity values associated to uniform hyperpriors are near zero, S(Un1,Un2) = 0.017, S(Un1,Un3) = 0.010, with calibrated values 0.017 and 0.010, despite the Hellinger distance between their corresponding priors being identical in gamma choices. In the case of the half-normal hyperpriors, the sensitivity associated to HN1 and HN2 is small (0.071 and calibrated value 0.069) but relevant when comparing HN1 and HN3 (S(HN1,HN3) = 0.588 and calibrated value 0.576).

4.4. Sensitivity of the risk of plot infection

The risk of plot infection was considered the most appropriate measure to describe results in Section 3 due to its great relevance in agronomic studies. In this sense, the analysis of the variability of the estimates from different modelling prior scenarios could be an important issue, mainly as a measure of confidence and reliability. As it was defined in (4), its posterior estimation will depend on the covariates, regression coefficients and random effects, which show different patterns regarding sensitivity. We carried out a sensitivity analysis for that on a similar basis as that for Section 3: the posterior distribution of the risk infection was calculated for a generic plot situated at altitude 76 meters (the sample median) for each virus and management conditions within each hyperprior scenario.

Table 3 shows the calibration of the Hellinger distance between the posterior distribution of the risk of plot infection for each management condition and virus. Similarly to the particular behaviour of the regression coefficients, the estimation of the risk of plot infection seems to be weakly influenced by the different hyperprior assumptions. In any case, the discrepancies observed between all the comparisons were not greater than the difference between the normal distribution $\text{N}(0,1)$ and $\text{N}(0.583,1)$, which reveals that Hellinger values are in general close to zero. It is worth noting that the Hellinger distance between normal distributions $\text{N}(0,1)$ and $\text{N}(1,1)$ is 0.343 (see again Table 4 in Appendix 2). In a similar manner, the uniform distributions had the smallest discrepancies together with half-normal distributions HN1 and HN2. However, as we expected inferences from HN3 exhibited the greatest discrepancies. Gamma hyperpriors showed substantial discrepancies, above all between Ga1 and Ga3, although these differences did not exceed those obtained from hyperprior HN3. Thus, these outcomes seem to indicate that the particular choice of a hyperprior distribution influences the estimation of the risk infection weakly but in a major extent that in the case of the estimates of the regression coefficients.

Table 3: *Calibration of the Hellinger distance between the posterior marginal distribution of the risk infection computed from hyperprior distributions Ga1 and Ga2, Ga1 and Ga3, Un1 and Un2, Un1 and Un3, HN1 and HN2, and HN1 and HN3.*

Virus	Management	(Ga1,Ga2)	(Ga1,Ga3)	(Un1,Un2)	(Un1,Un3)	(HN1,HN2)	(HN1,HN3)
ToMV	Organic	0.087	0.234	0.011	0.014	0.041	**0.583**
	Non-organic, greenhouse	0.051	0.139	0.011	0.011	0.029	0.355
	Non-organic, no greenhouse	0.041	0.100	0.015	0.016	0.031	0.268
CMV	Organic	0.079	0.213	0.015	0.014	0.041	0.536
	Non-organic, greenhouse	0.039	0.107	0.012	0.010	0.028	0.285
	Non-organic, no greenhouse	0.053	0.142	0.009	0.012	0.028	0.369
TSWV	Organic	0.049	0.128	0.026	0.025	0.037	0.323
	Non-organic, greenhouse	0.040	0.103	0.014	0.009	0.029	0.280
	Non-organic, no greenhouse	0.053	0.142	0.013	0.011	0.030	0.380

There are not so many discrepancies among the posterior means of the risk of a plot infection from the different hyperprior scenarios but there are many in the posterior variabilities (see Table 4). We accounted for variability in terms of standard deviation because it is a measure which describes the grade of uncertainty of the quantity of interest but mainly due to its direct agronomic interpretation. A great similarity in the posterior standard deviation values is repeatedly appreciated in results derived from Un1, Un2, Un3, HN2 and HN2 scenarios. The HN3 value was the most different. However, estimates corresponding to Ga1, Ga2 and Ga3 vary the most, especially in the case of Ga1.

Table 4: *Posterior standard deviation of the risk of a plot infection from the full inferential process based on Ga1, Ga2, Ga3, Un1, Un2, Un3, HN1, HN2 and HN3 hyperpriors.*

Virus	Management	Ga1	Ga2	Ga3	Un1	Un2	Un3	HN1	HN2	HN3
ToMV	Organic	0.136	0.146	0.161	0.184	0.184	0.184	0.183	0.178	0.118
	Non-organic, greenhouse	0.217	0.224	0.235	0.252	0.252	0.253	0.251	0.248	0.206
	Non-organic, no greenhouse	0.118	0.123	0.131	0.147	0.147	0.147	0.147	0.142	0.109
CMV	Organic	0.119	0.127	0.140	0.161	0.161	0.162	0.161	0.156	0.102
	Non-organic, greenhouse	0.161	0.166	0.175	0.190	0.190	0.190	0.189	0.186	0.151
	Non-organic, no greenhouse	0.179	0.186	0.198	0.216	0.216	0.216	0.215	0.211	0.166
TSWV	Organic	0.066	0.071	0.078	0.092	0.093	0.093	0.092	0.088	0.057
	Non-organic, greenhouse	0.172	0.178	0.187	0.203	0.202	0.202	0.201	0.198	0.162
	Non-organic, no greenhouse	0.185	0.192	0.204	0.223	0.223	0.224	0.222	0.218	0.172

In this sense, the posterior standard deviation for risk of a plot infection exhibits a considerable sensitivity to hyperparameter specification. For instance, the risk of a ToMV infection of a generic plot in an organic management system was estimated from 0.028 to 0.553 with 95% probability according to Ga1 scenario, but the subsequent interval in the Un1 scenario was [0.008,0.734].

5. Discussion

In this paper we have proposed a Bayesian correlated model (GLMM) to study and compare the risk of several virus infections in tomato and pepper plots under different agroecosystem conditions. First, we estimated several models, maintaining model specification but varying prior scenario default in accordance with different hyperprior distributions for the random effects scale parameter. Next, we conducted a sensitivity analysis to select the most stable model, in which effects of management conditions, altitude and random individual effects were assessed by estimating different derived quantities considered to be agronomically relevant.

Regarding the model covariates effect, the risk of plot infection was the quantity chosen to analyse agronomic outcomes. The risk of plot infection was estimated in the framework of mixed infections (with more than one virus) as well as in single infections (with only one virus). All the quantities applied for a “generic” plot of the population of each one of the agroecosystems considered. In the case of single infections, risk difference was also used to quantify differences among agroecosystems. Individual random effects were evaluated by assessing differences in the estimation of the risk of infection among plots managed under similar agroecosystem conditions. This enables the evaluation of the contribution of the common and of the individual elements in the model, and therefore the explanatory capacity of covariates.

In the case of mixed infections, organic agroecosystems exhibited lower prevalence for a three viruses joint infection. Non organic plots, independently of the presence of a greenhouse structure, showed a similar behaviour. Single infections were generally less prevalent or similar in organic systems than in conventional (non-organic with and without greenhouse), while TSWV and CMV infections were less prevalent under organic management; ToMV infection showed a slightly different behaviour pattern possibly as a consequence of the way it is transmitted (mechanical transmission). Altitude effect was clearly negative in all viruses but displayed considerable variability among the three viruses. Random effects behaviour was very regular in individuals belonging to non-organic with greenhouse and non-organic with no greenhouse considering that individual effects did not generate great differences among plots' risk infection estimates. Organic individuals exhibited more variable results in this aspect, but in general we can assume that all the fixed effects included in the model have a good explanatory capacity.

Sensitivity analysis was based on the methodology developed by Roos and Held (2011) and Roos et al. (2015). Hellinger distance and sensitivity measure, together with their corresponding calibration, allowed us to assess discrepancies in the estimation of the fixed effects (regression coefficients), the random effects standard deviation σ_b as well as the "generic" risk of infection among the prior scenarios tested. The evaluation of the posterior mean of the regression coefficients, the graphical characterization of the marginal posterior distribution of σ_b and the assessment of the standard deviation of the posterior distribution of the risk of plot infection among the several modelling scenarios completed the analysis. The outcomes obtained exhibited an insensitive behaviour of the fixed effects to hyperprior alterations with Hellinger values very close to zero and to each other. Only visual analysis of posterior means enabled us to detect a certain instability among inferences obtained from models under gamma hyperdistributions.

The estimation of σ_b showed a highly sensitive behaviour: gamma hyperpriors repeatedly exhibited the most relevant differences showing the greatest sensitivity values and the most divergent posterior distributions. In the case of risk infection estimation, in spite of all the Hellinger distances were around zero, gamma hyperdistributions showed interesting differences in terms of the standard deviation of the posterior distribution of the risk of plot infection. We therefore agree with Browne and Draper (2006), Roos et al. (2015), Roos and Held (2011), Gelman (2006), and Lunn et al. (2009) that gamma hyperpriors in hierarchical models lack robustness and a sensitivity analysis must be carried out in the Bayesian hierarchical framework to assess reliability of the performance. Furthermore, we also conclude that the "noninformative" nature of a hyperprior does not guarantee its impartiality in the inference process.

Appendix 1. The Hellinger distance

The Helliger distance (Le Cam, 2012) is a symmetric and invariant to any one-to-one transformation measure of discrepancy between two probability distributions, f and g, defined as follows

$$H(f,g) = \sqrt{\frac{1}{2}\int\limits_{-\infty}^{+\infty}\left(\sqrt{f(u)}-\sqrt{g(u)}\right)^2 \mathrm{d}u,}$$

where $0 \leq H(f,g) \leq 1$, 0 represents no divergence, and 1 full divergence.

The Hellinger distances between two gamma, uniform and half-truncated distributions are

- for gamma densities $\mathrm{Ga}(\alpha_1,\beta_1)$ and $\mathrm{Ga}(\alpha_2,\beta_2)$

$$H^2(\mathrm{Ga}(\alpha_1,\beta_1),\mathrm{Ga}(\alpha_2,\beta_2)) = 1-\Gamma\left(\frac{\alpha_1+\alpha_2}{2}\right)\sqrt{\frac{\beta_1^{\alpha_1}\beta_2^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)(\frac{\beta_1+\beta_2}{2})\alpha_1+\alpha_2}}$$

- for uniform densities $\mathrm{Un}(0,\eta_1)$ and $\mathrm{Un}(0,\eta_2)$, with $\eta_1 \leq \eta_2$

$$H^2(\mathrm{Un}(0,\eta_1),\mathrm{Un}(0,\eta_2)) = 1-\left(\frac{\eta_1}{\sqrt{\eta_1\eta_2}}\right)$$

- for half-normal densities $\mathrm{HN}(0,\sigma_1^2)$ and $\mathrm{HN}(0,\sigma_2^2)$

$$H^2(\mathrm{HN}(0,\sigma_1),\mathrm{HN}(0,\sigma_2)) = 1-\frac{\frac{1}{\sigma_1^2}\frac{1}{\sigma_2^2}^{1/4}}{\sqrt{\frac{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}}{2}}}$$

In the case of posterior distributions $\pi_1(\boldsymbol{\theta} \mid \mathscr{D})$ and $\pi_2(\boldsymbol{\theta} \mid \mathscr{D})$, the Hellinger distance can be approximated numerically at a finite set of K integration points as follows

$$H^2(\pi_1(\boldsymbol{\theta} \mid \mathscr{D}),\pi_2(\boldsymbol{\theta} \mid \mathscr{D})) = \frac{1}{2}\sum_{k=1}^{K}\left(\sqrt{\pi_1(\boldsymbol{\theta} \mid \mathscr{D})(k)}-\sqrt{\pi_2(\boldsymbol{\theta} \mid \mathscr{D})(k)}\right)^2 \Delta_k, \quad \text{(A.1)}$$

where the weights Δ_k are provided by the trapezoidal rule.

Appendix 2. Calibration

The Hellinger distance can be calibrated to evaluate the importance of the observed discrepancies by means of a reference parameter. Calibration was undertaken with respect to the normal distribution with variance one. The Hellinger distance between densities $\mathrm{N}(0,1)$ and $\mathrm{N}(\mu,1)$ is

$$H(\mathrm{N}(0,1),\mathrm{N}(\mu,1)) = \sqrt{1-\exp(-\mu^2/8)},$$

and consequently

$$\mu = \sqrt{-8\log(1-H^2(\mathrm{N}(0,1),\mathrm{N}(\mu,1)))}$$

Table A.2.1 shows a range of calibrated values μ with its subsequent Hellinger distance, $H(\mathrm{N}(0,1),\mathrm{N}(\mu,1))$.

Table A.2.1: *Calibration of the Hellinger distance.*

μ	$H(N(0,1),N(\mu,1))$
0	0
1	0.343
2	0.627
3	0.822
4	0.930
5	0.978
6	0.994
7	0.999
8	0.999
9	0.999
10	1

The sensitivity measure introduced previously can also be calibrated. Calibration of the sensitivity value obtained, s, has been obtained following the subsequent equation:

$$C(s,\mu') = \mu(s\times H(\mathrm{N}(0,1),\mathrm{N}(\mu',1))) \tag{A.2}$$

Interpretation of calibration can be conditioned by the choice of μ', so that for a value $\mu' = 1$, the value of s, would be comparable with the Hellinger distance obtained between two normal priors, $\mathrm{N}(0,1)$ and $\mathrm{N}(\mu' = 1,1)$ and the subsequent normal posteriors, $\mathrm{N}(0,1)$ and $\mathrm{N}(C(s,\mu' = 1),1)$. It is important to note that if $s > 1$ then $C(s,\mu') > \mu'$; if $s < 1$ then $C(s,\mu') < \mu'$; and if $s = 1$ then $C(s,\mu') = \mu'$.

Acknowledgements

We are very grateful to Josep Roselló, José Serra and Mª José Muñoz for providing the data and useful comments. We thank Marie Savage very much for reviewing the English and editing the manuscript. Elena Lázaro research was funded by the Spanish Ministry of Education, Culture and Sports, Grant FPU 2013/02042. The work of Carmen Armero was partially supported by Grant MTM2016-77501-P from the Spanish Ministry of Economy and Competitiveness. The work of Luis Rubio was partially funded by the INIA Grant RTA2013-00047-C02. We wish to acknowledge two anonymous referees and the Associate Editor for their valuable comments that substantially improved the original version of the paper.

References

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88, 669–679.

Alvares, D., Armero, C., Forte, C., and Rubio, L. (2016). Exploring Bayesian models to evaluate control procedures for plant disease. *Statistics and Operations Research Transactions, SORT*, 40, 139–152.

Bengtsson, J., Ahnström, J., and Weibull, A. C. (2005). The effects of organic agriculture on biodiversity and abundance: a meta-analysis. *Journal of Applied Ecology*, 42, 261–269.

Bettiol, W., Ghini, R., Galvão, J. A. H., and Siloto, R. C. (2004). Organic and conventional tomato cropping systems. *Scientia Agricola*, 61, 253–259.

Box, G. E., and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. Hoboken: John Wiley & Sons.

Browne, W. J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1, 473–514.

Clark, M. F., Adams, A. N. and Barbara, D. J. (1976). The detection of plant viruses by enzyme-linked immunosorbent assay (ELISA). In *X International Symposium on Fruit Tree Virus Diseases* 67, 43–50.

Clark, J. S., Wolosin, M., Dietze, M., Ibanez, I., Ladeau, S., Welsh, M. and Kleoppel, B. (2007). Tree growth inference and prediction from diameter censuses and ring widths. *Ecological Applications*, 17, 1942–1953.

Christensen, R., Johnson, W., Branscum, A. and Hanson, T. E. (2011). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Boca Raton: Chapman & Hall/CRC Press.

Finley, A. O., Banerjee, S., and Basso, B. (2011). Improving Crop Model Inference Through Bayesian Melding with Spatially-Varying Parameters. *Journal of Agricultural, Biological and Environmental Statistics*, 16, 453–474.

Gallitelli, D. (2000). The ecology of Cucumber mosaic virus and sustainable agriculture. *Virus Research*, 71, 9–21.

García-Cano, E., Resende, R. O., Fernández-Muñoz, R. and Moriones, E. (2006). Synergistic interaction between Tomato chlorosis virus and Tomato spotted wilt virus results in breakdown of resistance in tomato. *Phytopathology*, 96, 1263–1269.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1, 515–534.

Gelman, A. and Rubin, D.B. (1992). Inference from Iterative Simulation using Multiple Sequences. *Statistical Science*, 7, 457–511.

Hanssen, I. M., Lapidot, M. and Thomma, B. P. (2010). Emerging viral diseases of tomato crops. *Molecular Plant-Microbe Interactions*, 23, 539–548.

Jeffreys, H. (1998). *The Theory of Probability*. Third edition. New York: Oxford University Press.

Krachey, M. and Boone, E.L. (2012). bmk: MCMC diagnostics package. R package version 1.0. *http://CRAN.R-project.org/package=bmk*.

Lambert, P.C., Sutton, A.J., Burton, P.R., Abrams, K.R. and Jones, D.R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24, 2401–2428.

Le Cam, L. (2012). *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag.

Letourneau, D.K. and Goldstein, B. (2001). Pest damage and arthropod community structure in organic vs. conventional tomato production in California. *Journal of Applied Ecology*, 38, 557–570.

Loredo, T.J. (1990). From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics (1990). In *Maximum Entropy and Bayesian Methods* (P.F. Fougère eds), 81-142. Dordrecht: Kluwer Academic.

Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2012). *The BUGS Book: A practical Introduction to Bayesian Analysis*. Boca Raton: Chapman & Hall/CRC Press.

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: Concepts, Structure and Extensibility. *Statistics and Computing*, 10, 325–337.

Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009). The BUGS Project: Evolution, Critique and Future Directions. *Statistics in Medicine*, 28, 3049–3067.

McCulloch, R.E. (1989). Local Model Influence. *Journal of the American Statistical Association*, 84, 473–478.

Murphy, J.F. and Bowen, K.L. (2006). Synergistic disease in pepper caused by the mixed infection of Cucumber mosaic virus and Pepper mottle virus. *Phytopathology*, 96, 240–247.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. Hoboken: John Wiley & Sons.

Paciorek, C.J. and McLachlan, J.S. (2009). Mapping Ancient Forests: Bayesian Inference for Spatio-Temporal Trends in Forest Composition Using the Fossil Pollen Proxy Record. *Journal of the American Statistical Association*, 104, 608–622.

Paradinas, I., Conesa, D., Pennino, MG., Muñoz, F., Fernández, AM., López-Quílez, A. and Bellido, JM. (2015). Bayesian spatio-temporal approach to identifying fish nurseries by validating persistence areas. *Marine Ecology Progress Series*, 52, 245–255.

Roos, M. and Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, 6, 259–278.

Roos, M., Martins, T. G., Held, L. and Rue, H. (2015). Sensitivity analysis for Bayesian hierarchical models. *Bayesian Analysis*, 10, 321–349.

Serra, J., Ocon, C., Jiménez, A., Arnau, J., Malagón, J. and Porcuna, J.L. (1999). Epidemiología de las virosis en la Comunidad Valenciana: el caso del "virus de la cuchara" del tomate. *Comunidad Valenciana Agraria*, 14, 47–53.

Spiegelhalter, D.J., Abrams, K.R. and Myles, J.P. (2004). *Bayesian Approaches to Clinical trials and Health-Care evaluation*. Chichester: John Wiley & Sons.

Thompson, S.G., Smith, T.C. and Sharp, S.J. (1997). Investigating underlying risk as a source of heterogeneity in meta-analysis. *Statistics in Medicine*, 16, 2741–2758.

Thornley, J.H. and France, J. (2007). *Mathematical Models in Agriculture: Quantitative Methods for the Plant, Animal and Ecological Sciences*. Oxon: Cabi.

Tomlinson, J.A. (1987). Epidemiology and control of virus diseases of vegetables. *Annals of Applied Biology*, 110, 661–681.

Van Bruggen, A.H. (1995). Plant disease severity in high-input compared to reduced-input and organic farming systems. *Plant Disease*, 79, 976–984.

SORT 41 (1) January-June 2017, 117-118 DOI: 10.2436/20.8080.02.53

Corrigendum to “Transmuted geometric distribution with applications in modelling and regression analysis of count data”

Subrata Chakraborty[1] and Deepesh Bhati[2]

In our paper “Transmuted geometric distribution with applications in modelling and regression analysis of count data” (Chakraborty and Bhati, 2016) there is a mistake in the expression for the variance $\mathbb{V}(Y)$, which induced a wrong expression for the Index of Dispersion (ID) in Table 1, a wrong figure in Figure 2(c) and a wrong sentence in the paragraph just preceding Remark 5. These were noticed only after the article was published. These corrections, as well as few corrections in Table 6 and a correction of one reference, are as follows:

i. In Table 1 S.No. 2 the correct expression for $\mathbb{V}(Y)$ should be read as,

$$\frac{q\left(1-\alpha+q\left(2+q(1-\alpha)-\alpha^2\right)\right)}{\left(1-q^2\right)^2}$$

ii. In Table 1 S.No. 3 the correct expression for Index of Dispersion (ID) should be read as,

$$\frac{1-\alpha+q\left(2+q(1-\alpha)-\alpha^2\right)}{\left(1-q^2\right)\left(1+q-\alpha\right)}$$

iii. Figure 2(c) should be seen as,

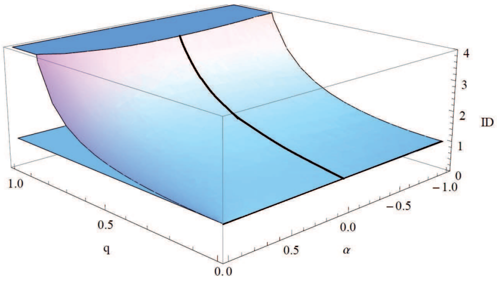


Figure 2(c): *$q-\alpha$ surface plot of Index of Dispersion of $\mathscr{TGD}(q,\alpha)$*

[1] Department of Statistics, Dibrugarh University, Dibrugarh-786004, Assam, India.

[2] Department of Statistics, Central University of Rajasthan, Ajmer-305817, Rajasthan, India.

Received: September 2016
Accepted: November 2016

iv. Finally in pages 161-162 the sentence “Further, Figure 2(c) shows that the horizontal q-α surface cuts the ID surface at 1 indicating under or over-dispersion for $\alpha \in (1,0)$ or $(0,1)$ respectively (see Remark 3)” should be corrected as “Further, Figure 2(c) shows that the horizontal q-α surface is always above 1 indicating over-dispersion for all the values of α”. In Table 6: column 5, row 2, -0.983 should be read as 0.0; column 6, row 2, 0.2228 should be read as 1.255; column 5, row 9, -0.1160 should be read as -0.1559.

v. In the references, page 175, publication year of Sastry et al.’s reference should be read as 2014.

Both authors apologize for these errors.

References

Chakraborty, S., and Bhati, D. (2016). Transmuted geometric distribution with applications in modelling and regression analysis of count data. *SORT-Statistics and Operations Research Transactions*, 40(1), 153–176. URL: http://www.idescat.cat/sort/sort401/40.1.7.chakraborty-bhati.pdf
DOI: 10.2436/20.8080.02.39

SORT 41 (1) January-June 2017, 119-138 DOI: 10.2436/20.8080.02.54

Goodness-of-fit test for randomly censored data based on maximum correlation

Ewa Strzalkowska-Kominiak[1] and Aurea Grané[2]

Abstract

In this paper we study a goodness-of-fit test based on the maximum correlation coefficient, in the context of randomly censored data. We construct a new test statistic under general right-censoring and prove its asymptotic properties. Additionally, we study a special case, when the censoring mechanism follows the well-known Koziol-Green model. We present an extensive simulation study on the empirical power of these two versions of the test statistic, showing their advantages over the widely used Pearson-type test. Finally, we apply our test to the head-and-neck cancer data.

MSC: 62N01, 62N03, 62G10, 62G20.

Keywords: Goodness-of-fit, Kaplan-Meier estimator, maximum correlation, random censoring.

1. Introduction

In many medical studies one encounters data which are not fully observed but censored from the right. For example, in the head-and-neck cancer trial studied by Nikulin and Haghighi (2006), one observes survival times for 42 out of 51 patients, whereas for the remaining 9 patients only the time to follow-up is given. Let $Y_1,\ldots,Y_n$ be the lifetimes of interest, e.g., the survival times of head-and-neck cancer patients, coming from a continuous distribution function F and let $C_1,\ldots,C_n$ be the censoring times (that is, the times to follow-up) coming from a distribution function G. In the context of right-censored data, for every $i=1,\ldots,n$, we observe

$$X_i = \min(Y_i, C_i) \text{ and } \delta_i = 1_{\{Y_i \le C_i\}},$$

[1] Mathematical Institute, Cracow University of Technology, Warszawska 24, 31-155 Kraków, Poland. estrzalkowska-kominiak@pk.edu.pl

[2] Statistics Department, Universidad Carlos III de Madrid, c/ Madrid 126, 28903 Getafe, Spain. aurea.grane@uc3m.es (corresponding author).

Received: July 2016
Accepted: December 2016

where 1_A denotes the indicator function, being equal to 1 if A is fulfilled and 0 otherwise. The unknown distribution function of the lifetimes F can be estimated by the well-known product-limit estimator introduced by Kaplan and Meier (1958). However, if the shape of the distribution could be assumed, there would be a substantial gain in the efficiency of statistical procedures. For instance, in the example of head-and-neck cancer data, Nikulin and Haghighi (2006) suggest that the lifetimes follow the Generalized-Power Weibull family and that hypothesis is tested. Therefore, goodness-of-fit tests are an important statistical tool when dealing with (right-)censored data. Under complete data set-up we have a multitude of goodness-of-fit tests to select from. See, e.g., Darling (1957) or Massey (1951) for the historical literature on the subject and Torabi et al. (2016) or Novoa-Muñoz and Jiménez-Gamero (2016), among many others, for the most recent publications. Some widely used tests for complete data, like Kolmogorov-Smirnov or Cramer-von Mises, are difficult to apply in the presence of censoring, since the limit distribution depends on the censoring distribution G. See Balakrishnan et al. (2015) for a recent overview on this kind of tests with randomly censored data. Other classical approaches are Koziol and Green (1976) and Akritas (1988). The former is more restrictive, since it is based on the assumption that the distribution function G follows the so called Koziol-Green model, whereas the latter is a χ^2 test applied to general random censoring. This is the reason why the Pearson-type goodness-of-fit test proposed by Akritas (1988) is so far the best option for randomly censored data with unknown censoring distribution. Nevertheless, it requires a partition of the observations into cells jointly with an adequate choice of number of classes, since the power of the test may vary depending on the degrees of freedom. In this work we propose a new goodness-of-fit test based on the maximum correlation coefficient, with normal limiting distribution and, therefore, straightforward to apply.

We start by introducing the maximum correlation in a more general set-up. Let Y_1 and Y_2 be two random variables with finite second order moments, joint cumulative distribution function (cdf) H and marginals F_1 and F_2, respectively. The Hoeffding representation of the correlation coefficient is given by

$$\rho(F_1,F_2) = \frac{1}{\sigma_1\sigma_2}\int_{\mathbb{R}^2}(H(x,y) - F_1(x)F_2(y))dxdy,$$

where σ_i denotes the standard deviation of Y_i. Furthermore, the maximum correlation of the pair of random variables (Y_1,Y_2) is defined as the correlation coefficient $\rho^+(F_1,F_2)$ corresponding to the bivariate cdf $H^+(x,y) = \min(F_1(x),F_2(y))$, the upper Fréchet bound of $H(x,y)$. The cdf $H^+(x,y)$ is a singular distribution, having support on the one-dimensional set $\{(x,y) \in \mathbb{R}^2 : F_1(x) = F_2(y)\}$, and the maximum correlation coefficient is given by

$$\rho^+(F_1,F_2) = \frac{1}{\sigma_1\sigma_2}\left(\int_0^1 F_1^{-1}(p)F_2^{-1}(p)dp - \mu_1\mu_2\right), \tag{1}$$

where F_i^{-1} is the inverse of F_i and μ_i is the mean of Y_i. This maximum correlation, $\rho^+(F_1,F_2)$, is a measure of agreement between F_1 and F_2, since $\rho^+ = 1$ if and only if $F_1 = F_2$ up to a scale and location change. In particular, Cuadras and Fortiana (1993) proposed the statistic based on $\rho^+(F,F_0)$ as a measure of goodness of fit of an iid sample $Y_1,\dots,Y_n$ with cdf F, to a given distribution F_0. The goodness-of-fit test based on maximum correlation was further studied by Fortiana and Grané (2003), Grané (2012) and Grané and Tchirina (2013).

As in the latter publications, the present paper is devoted to testing uniformity, i.e. $F_0 = F_U$, a $[0,1]$ uniform distribution. As shown by Fortiana and Grané (2003) the asymptotic approximation of $\rho^+(F,F_U)$ is available, but convergence to its limiting law is rather slow. This led to defining

$$Q = \frac{\sigma}{\sqrt{1/12}}\rho^+(F,F_U) = 6\int_0^1 x(2F(x)-1)F(dx), \tag{2}$$

where σ is the standard deviation of $Y_i \sim F$, which equals one if $F = F_U$.

The goal of this paper is to study a test statistic based on Q when $Y_1,\dots,Y_n$ may not be fully observed but censored from the right by censoring times $C_1,\dots,C_n$. More precisely, we wish to test the hypothesis $H_0 : F = F_U$, where F_U is the cdf of a $[0,1]$ uniform random variable, based on the sample $(X_i,\delta_i)_{i=1,\dots,n}$, where $X_i = \min(Y_i,C_i)$, with $X_i \in [0,1]$. Nevertheless, our approach is not restricted to testing uniformity. We can also consider a more general null hypothesis F_0, since the transformed random variable $F_0(Y)$ follows a $[0,1]$ uniform distribution under $H_0 : F = F_0$. That is, $\tilde{Y} = F_0(Y) \sim F_U$ under the null hypothesis. Then, setting $\tilde{C} = F_0(C)$ and since $\{\tilde{Y}_i \le \tilde{C}_i\} = \{Y_i \le C_i\}$, leads us to testing uniformity based on the iid sample $(\tilde{X}_1,\delta_1),\dots,(\tilde{X}_n,\delta_n)$, where

$$\tilde{X}_i = \min(\tilde{Y}_i,\tilde{C}_i) \text{ and } \delta_i = 1_{\{\tilde{Y}_i \le \tilde{C}_i\}}.$$

Hence, testing for uniformity is equivalent to testing for a fully specified continuous distribution. Even though it seems that we could extend the work of Fortiana and Grané (2003) by setting $Q_n = 6\int_0^1 x(2F_n(x)-1)F_n(dx)$, where F_n denotes the Kaplan-Meier estimator for censored data, it is far from being true. In contrast to the empirical distribution under completely observed data, the Kaplan-Meier estimator is biased (see Stute (1994), for details). In Section 2 we show that such a plug-in estimator suffers from the bias of the product-limit estimator and, therefore, $E(Q_n) = 1$ does not hold under H_0. To avoid this problem we propose to re-write Q in such a way that it can be estimated by U-statistics. This leads to significant bias (and variance) reduction. In Section 3 we prove the asymptotic normality of the proposed estimator and in Section 4 we present our new goodness-of-fit test. In Section 5 we present an extensive simulation study. Finally, in Section 6 we adapt the test statistic to the case of composite null hypothesis and apply our test to the head-and-neck cancer data from Nikulin and Haghighi (2006).

2. Test statistic

In this section we propose our new goodness-of-fit statistic for randomly censored data, based on the modified maximum correlation coefficient. Recall that, under $H_0 : F = F_U$, the quantity

$$Q = \frac{\sigma}{\sqrt{1/12}}\rho^+(F,F_U) = 6\int_0^1 x(2F(x)-1)F(dx)$$

equals one. Hence in the following we prefer to work with

$$Q^1 = Q - 1 = 6\int_0^1 x(2F(x)-1)F(dx) - 1 \tag{3}$$

which equals zero if H_0 is true.

First, we define a plug-in estimator of Q^1 by replacing F in (3) with the well-known Kaplan-Meier estimator. We obtain

$$Q_n^1 = 6\int_0^1 x(2F_n(x)-1)F_n(dx) - 1, \tag{4}$$

where F_n is defined as follows

$$F_n(x) = 1 - \prod_{X_i \le x}\left[1 - \frac{\delta_i}{\sum_{k=1}^n 1_{\{X_k \ge X_i\}}}\right]. \tag{5}$$

It turns out that, under the null hypothesis and for finite samples, the plug-in estimator Q_n^1 suffers from significant bias and its convergence to the limiting distribution is very slow.

To solve this problem, we propose to estimate Q^1 with a U-statistic. For this, note that if F is a continuous cdf and $supp(F) \subseteq [0,1]$, then

$$2\int_0^1 F(x)F(dx) = 1.$$

Hence

$$Q^1 = \int_0^1 (6x(2F(x)-1) - 2F(x))F(dx) = \int_0^1 [(6x-2)F(x) - 6x(1-F(x))]F(dx)$$

$$= \int_0^1\int_0^1 [(6x-2)1_{\{y\le x\}} - 6x1_{\{y>x\}}]F(dx)F(dy). \tag{6}$$

Now we may replace the unknown quantities by their estimators. For this we introduce the jumps of the Kaplan-Meier estimator by setting

$$w_{in} = F_n(X_i) - F_n(X_i-),$$

where $F_n(x-)$ is the left-continuous version of $F_n(x)$, which is defined analogously as (5) but with the product over all $X_i < x$.

Finally, the estimator of Q^1 is given by

$$\tilde{Q}_n = \sum_{i=1}^{n} \sum_{j \neq i} w_{in} w_{jn} h(X_i, X_j), \tag{7}$$

where

$$h(x_1, x_2) = (6x_1 - 2) 1_{\{x_2 \leq x_1\}} - 6x_1 1_{\{x_2 > x_1\}}.$$

To illustrate the advantages of using $\tilde{Q}_n$ over the plug-in estimator Q_n^1, in panel (a) of Figure 1, we present the bias and variance of those estimators under the null hypothesis, that is, when the data come from the $[0,1]$ uniform distribution. Additionally, in panels (b)-(c) of Figure 1, we compare the kernel density estimators of the standardized versions of $\tilde{Q}_n$ and Q_n^1 to that of the standard normal distribution. The standardization is done using the estimated asymptotic variances, discussed later on. Clearly, the U-statistic $\tilde{Q}_n$ exhibits much smaller bias (and variance) than Q_n^1 and, additionally, its standardized version fits nicely the standard normal distribution for all the considered censoring rates.

(a) Estimated bias (variance)

10% censoring

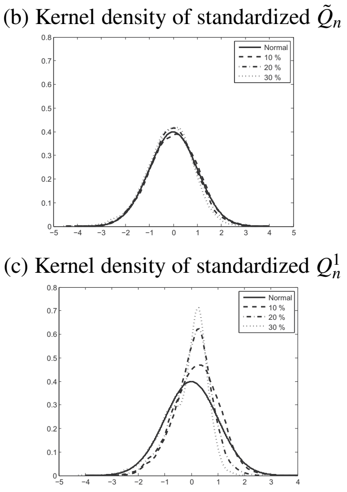
	Q_n^1	$\tilde{Q}_n$
$n = 50$	0.0338 (0.0081)	-0.0010 (0.0047)
$n = 100$	0.0150 (0.0032)	-0.0014 (0.0022)

20% censoring

	Q_n^1	$\tilde{Q}_n$
$n = 50$	0.0111 (0.0197)	-0.0069 (0.0063)
$n = 100$	0.0047 (0.0074)	-0.0026 (0.0026)

30% censoring

	Q_n^1	$\tilde{Q}_n$
$n = 50$	-0.0206 (0.0483)	-0.0122 (0.0069)
$n = 100$	-0.0275 (0.0204)	-0.0090 (0.0033)


(b) Kernel density of standardized $\tilde{Q}_n$

(c) Kernel density of standardized Q_n^1

Figure 1: *Comparison between $\tilde{Q}_n$ and the plug-in estimator Q_n^1: Estimated bias (variance) based on* 5000 *trials and kernel densities for $n = 200$.*

3. Asymptotic properties

In this section we study the asymptotic properties of our test statistic $\tilde{Q}_n$. Firstly, we consider $\tilde{Q}_n$ under a general censoring mechanism, that is, without assuming any shape for the distribution function of the censoring times $G(x) = P(C \leq x)$. Secondly, we apply the results to the special case of the Koziol-Green model. Recall that $F(x) = P(Y \leq x)$ is the cdf of the lifetimes of interest. We need the following assumptions A1-A2, which assure that the asymptotic variance is bounded and censoring is not too heavy. These conditions allow us to apply the limit theorems from Stute (1995) in order to prove the asymptotic normality:

$$A1 : \int_0^1 \frac{F(du)}{1-G(u)} < \infty$$

$$A2 : \int_0^1 |\varphi(u)| C^{1/2}(u) F(du) < \infty$$

where $\varphi(x) = 12xF(x) - 6x - 2 - 12\int_0^x yF(dy) + 6\int_0^1 yF(dy)$ is a score function, $C(x) = \int_0^x \frac{G(dy)}{(1-G(y))^2(1-F(y))}$ and F is continuous with support in $[0,1]$.

Theorem 1 *Under A1 and A2, we have*

$$\sqrt{n}(\tilde{Q}_n - Q^1) \to \mathcal{N}(0, \sigma^2),$$

where

$$\sigma^2 = \int_0^1 \frac{\varphi^2(x)}{1-G(x)} F(dx) - \left[\int_0^1 \varphi(x)F(dx)\right]^2 - \int_0^1 \left[\int_x^1 \varphi(y)F(dy)\right]^2 \frac{(1-F(x))G(dx)}{(1-H(x))^2}$$

and

$$\varphi(x) = 12xF(x) - 6x - 2 - 12\int_0^x yF(dy) + 6\int_0^1 yF(dy).$$

Proof. See Appendix.

Consequently, we have that

Corollary 1 *Under H_0, A1 and A2, we have*

$$\sqrt{n}\tilde{Q}_n \to \mathcal{N}(0, \sigma^2).$$

The variance under H_0 would not simplify, since it does depend on the distribution function of the censoring times G, which is unknown. Nevertheless, under the Koziol-

Green model, we have an explicit expression for σ^2. First, recall that G follows a Koziol-Green model if

$$1 - G(x) = (1 - F(x))^{\beta},$$

where $\beta > 0$ is an unknown parameter. However, we can see that

$$p := P(Y > C) = \frac{\beta}{\beta + 1} \text{ and } 1 - p = \int (1 - G(x))F(dx).$$

Hence β can be easily estimated using Kaplan-Meier estimators for F and G. Finally, it is easy to check that assumptions A1 and A2 are fulfilled under the Koziol-Green model with $\beta \in (0,1)$, that is, if the censoring is not heavier than 50%, which is a very reasonable assumption. So, as a consequence of Corollary 1, we get the following result.

Corollary 2 *Under the Koziol-Green model with $\beta \in (0,1)$ we have that, under H_0,*

$$\sqrt{n}\tilde{Q}_n \to \mathcal{N}(0, \sigma^2_{KG}),$$

where

$$\sigma^2_{KG} = \frac{-\beta^4 + 4\beta^3 - 17\beta^2 + 38\beta - 24}{(\beta-1)(\beta-2)(\beta-3)(\beta-4)(\beta-5)}.$$

4. Goodness-of-fit test

Once the test statistic is proposed and its limiting distribution is established, we are in the position to define the goodness-of-fit test. For this we estimate the asymptotic variance σ^2 using the plug-in principle, that is, by replacing the unknown quantities with their estimators. First, we define the distribution function of the observed times $\tilde{H}(x) = P(X \leq x)$ and set $\tilde{H}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}$ as its empirical counterpart. Moreover, let

$$H^0(x) = \mathbb{P}(X \leq x, \delta = 0) = \int_0^x (1 - F(u))G(du)$$

and

$$H^1(x) = \mathbb{P}(X \leq x, \delta = 1) = \int_0^x (1 - G(u))F(du)$$

be the subdistributions related to the observed censored and uncensored lifetimes. Their estimators are defined as follows

$$H^0_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}(1 - \delta_i)$$

and

$$H_n^1(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \le x\}}\delta_i.$$

Hence

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\varphi_n^2(X_i)}{(1-G_n(X_i-))^2}\delta_i - \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\varphi_n(X_i)}{1-G_n(X_i-)}\delta_i\right]^2$$

$$-\frac{1}{n}\sum_{i=1}^{n} \frac{1-\delta_i}{(1-H_n(X_i-))^2}\left[\frac{1}{n}\sum_{j=1}^{n} \frac{\varphi_n(X_j)}{1-G_n(X_j-)}\delta_j 1_{\{X_j \ge X_i\}}\right]^2,$$

where

$$\varphi_n(x) = 12xF_n(x) - 6x - 2 - 12\frac{1}{n}\sum_{i=1}^{n} \frac{X_i\delta_i}{1-G_n(X_i-)}1_{\{X_i \le x\}} + 6\frac{1}{n}\sum_{i=1}^{n} \frac{X_i\delta_i}{1-G_n(X_i-)}.$$

and G_n is a Kaplan-Meier estimator given by

$$1 - G_n(x) = \prod_{X_i \le x}\left[1 - \frac{1-\delta_i}{\sum_{k=1}^{n} 1_{\{X_k \ge X_i\}}}\right].$$

Before we may define the goodness-of-fit test, we need to show the consistency of the variance estimator σ_n^2. For this, we require an assumption which is stronger than A1. In particular:

$A3$: There exists $\varepsilon > 0$ such that $\displaystyle\int_0^1 \frac{F(dx)}{(1-G(x))^{1+\varepsilon}} < \infty$

Lemma 1 *Under A3, we have*

$$\sigma_n^2 \xrightarrow{P} \sigma^2$$

Proof. See Appendix.

Finally, we have

Theorem 2 *Under H_0 and assumptions A2 and A3, we have that*

$$T_n := \frac{\sqrt{n}\tilde{Q}_n}{\sqrt{\sigma_n^2}} \xrightarrow{d} N(0,1). \tag{8}$$

Proof. The result follows from Corollary 1 and Lemma 1.This completes the proof. ∎

In view of Theorem 1, we reject H_0 at level α if

$$T_n \leq \Phi^{-1}(\alpha/2) \text{ or } T_n \geq \Phi^{-1}(1-\alpha/2),$$

where Φ^{-1} is the inverse of the standard normal cdf.

Additionally, under the Koziol-Green model and in view of Corollary 2, we define

$$T_n^{KG} := \frac{\sqrt{n}\tilde{Q}_n}{\sqrt{\hat{\sigma}^2_{KG}}}, \tag{9}$$

where

$$\hat{\sigma}^2_{KG} = \frac{-\hat{\beta}^4 + 4\hat{\beta}^3 - 17\hat{\beta}^2 + 38\hat{\beta} - 24}{(\hat{\beta}-1)(\hat{\beta}-2)(\hat{\beta}-3)(\hat{\beta}-4)(\hat{\beta}-5)}$$

and

$$\hat{\beta} = \left(\int (1-G_n(x))F_n(dx)\right)^{-1} - 1.$$

It is easy to see that

$$\int (1-G_n(x))F_n(dx) \xrightarrow{P} \int (1-G(x))F(dx).$$

Hence $\hat{\beta} \xrightarrow{P} \beta$ and $\hat{\sigma}^2_{KG} \xrightarrow{P} \sigma^2$. Consequently, as before, we reject H_0 at level α if

$$T_n^{KG} \leq \Phi^{-1}(\alpha/2) \text{ or } T_n^{KG} \geq \Phi^{-1}(1-\alpha/2).$$

5. Simulation study

Here we conduct an extensive simulation study to show the behaviour of our test. In the following subsection we consider only the null hypothesis, while in Subsection 5.2 we include the power study under different families of alternatives. In both subsections we compare our method with the Pearson-type goodness-of-fit test proposed by Akritas (1988). Following the notation of Section 4, we denote by T_n and T_n^{KG} our test statistics for the general censoring and under the Koziol-Green model, respectively. See, equations (8) and (9) for details. Moreover, we denote by $A_{(nc)}$ the χ^2 test proposed by Akritas (1988), where nc denotes the number of cells.

5.1. Null hypothesis

In this section we present the results of the proposed methods under the null hypothesis and at 5% significance level. As mentioned before, we consider our tests T_n and T_n^{KG}, together with the test presented by Akritas (1988). Following the latter work, we consider $nc = 2$ and $nc = 5$ and denote these tests by $A_{(2)}$ and $A_{(5)}$, respectively. The results are based on 5000 trials. From Table 1 we see that tests T_n and those from Akritas hold very well the significance level. The test based on the Koziol-Green model holds the 5% level when censoring is low. However, for more than 20% of missing data, the variance σ^2_{KG} does not captures the variability of our $\tilde{Q}_n$ correctly and, therefore, the significance level is slightly overestimated for heavy censoring.

Table 1: *Empirical level for testing null hypothesis.*

	10% censoring				20% censoring				30% censoring			
	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}
$n = 50$	0.0508	0.0552	0.0538	0.0578	0.0450	0.0560	0.0558	0.0530	0.0496	0.0596	0.0560	0.0648
$n = 100$	0.0480	0.0582	0.0548	0.0532	0.0478	0.0530	0.0492	0.0560	0.0502	0.0568	0.0528	0.0708
$n = 200$	0.0468	0.0524	0.0574	0.0498	0.0522	0.0508	0.0508	0.0604	0.0494	0.0522	0.0468	0.0666

5.2. Power study

In order to study the power of our test we consider two different models:

Model 1: To test the uniformity ($H_0 : F = F_U$) we choose three parametric families of alternative probability distributions with support on [0, 1]:

(a) Lehmann alternatives,

$$F_\theta(x) = x^\theta, 0 \leq x \leq 1, \theta \geq 1;$$

where for $\theta = 1$ we have $F_\theta = F_U$.

(b) compressed uniform alternatives,

$$F_\theta(x) = \frac{x-\theta}{1-2\theta}, \ \theta \leq x \leq 1-\theta,$$

where $0 \leq \theta \leq 1/2$; and for $\theta = 0$ we have $F_\theta = F_U$.

(c) centred distributions having a U-shaped density for $\theta \in (0,1)$, or wedge-shaped density for $\theta > 1$

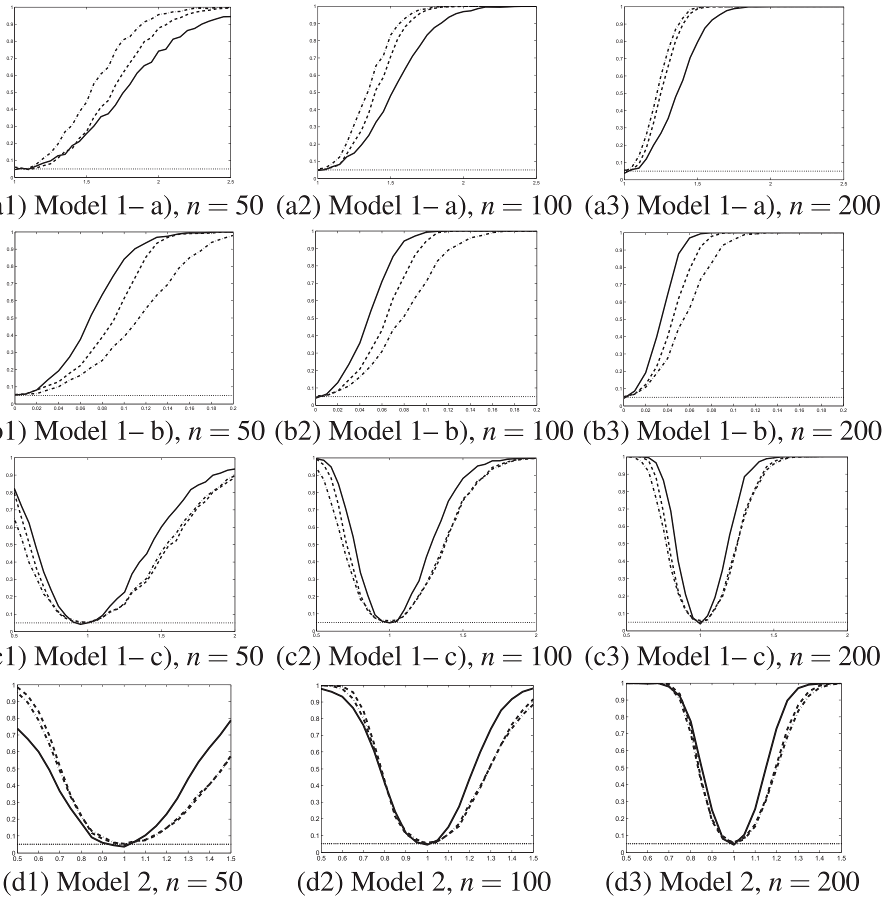
$$F_\theta(x) = \begin{cases} \frac{1}{2}(2x)^\theta, & 0 \leq x \leq 1/2 \\ 1 - \frac{1}{2}(2(1-x))^\theta, & 1/2 \leq x \leq 1 \end{cases}$$

where for $\theta = 1$ we have $F_\theta = F_U$.

Model 2: An exponentiality test (with parameter $\lambda = 1$), where the alternatives are Weibull distributions with parameters 1 and θ. More precisely, $F_\theta(x) = 1 - e^{-x^\theta}$, where $\theta = 1$ gives us the exponential distribution of the null hypothesis.

Additionally, the censoring variable C is generated under the Koziol-Green model. That is, $1 - G(x) = (1 - F(x))^\beta$, where $\beta = \frac{p}{1-p}$ and $p = \mathbb{P}(X > C)$ is the censoring level.

In the following figures and tables we present the power study at a 5% significance level. Panels (a1)-(c3) of Figure 2 contain the power of the test for Model 1 and panels (d1)-(d3) of Figure 2 contain the power under Model 2, for different sample sizes ($n = 50, 100, 200$) and one censoring level of 20%. All those figures are based on 2000 trials.



(a1) Model 1– a), $n = 50$ (a2) Model 1– a), $n = 100$ (a3) Model 1– a), $n = 200$

(b1) Model 1– b), $n = 50$ (b2) Model 1– b), $n = 100$ (b3) Model 1– b), $n = 200$

(c1) Model 1– c), $n = 50$ (c2) Model 1– c), $n = 100$ (c3) Model 1– c), $n = 200$

(d1) Model 2, $n = 50$ (d2) Model 2, $n = 100$ (d3) Model 2, $n = 200$

Figure 2: *Power study for Model 1 (a1–c3) and Model 2 (d1–d3) for three different sample sizes and censoring rate $p = 0.2$. T_n (solid line), $A_{(5)}$ (dashed line) and $A_{(2)}$ (dash-dotted line).*

Table 2: *Power study for Model 1 and Model 2.*

Model 1, Alternative a)												
$n = 100$	$p = 0.1$				$p = 0.2$				$p = 0.3$			
	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}
$\theta = 1.5$	0.5277	0.7107	0.8040	0.5547	0.4760	0.7220	0.8190	0.5167	0.3957	0.7093	0.8260	0.4570
$\theta = 2$	0.9847	0.9997	0.9997	0.9873	0.9997	1	1	1	0.9070	0.9980	0.9995	0.9520
$\theta = 2.5$	1	1	1	1	0.9997	1	1	1	0.9237	0.9990	0.9997	0.9607
$n = 200$	$p = 0.1$				$p = 0.2$				$p = 0.3$			
	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}
$\theta = 1.5$	0.8410	0.9823	0.9887	0.8540	0.8020	0.9830	0.9893	0.8200	0.7057	0.9833	0.9910	0.7517
$\theta = 2$	0.9997	1	1	0.9997	1	1	1	1	0.9950	1	1	0.9990
$\theta = 2.5$	1	1	1	1	1	1	1	1	0.9983	1	1	0.9997
Model 1, Alternative b)												
$n = 100$	$p = 0.1$				$p = 0.2$				$p = 0.3$			
	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}
$\theta = 0.05$	0.5987	0.3897	0.2940	0.5473	0.5370	0.3027	0.2297	0.5440	0.4563	0.2530	0.1803	0.5037
$\theta = 0.1$	1	0.9837	0.8510	0.9987	0.9957	0.9413	0.7163	0.9950	0.9410	0.9000	0.5910	0.9820
$\theta = 0.15$	1	1	0.9963	1	1	1	0.9797	1	0.9940	1	0.9500	1
$n = 200$	$p = 0.1$				$p = 0.2$				$p = 0.3$			
	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}
$\theta = 0.05$	0.9133	0.7517	0.5820	0.8830	0.8560	0.6003	0.4250	0.8313	0.7683	0.4870	0.3287	0.7973
$\theta = 0.1$	1	1	0.9933	1	1	1	0.9570	1	0.9977	0.9987	0.8977	1
$\theta = 0.15$	1	1	1	1	1	1	1	1	1	1	0.9993	1
Model 1, Alternative c)												
$n = 100$	$p = 0.1$				$p = 0.2$				$p = 0.3$			
	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}
$\theta = 0.75$	0.6587	0.3623	0.3217	0.6803	0.5337	0.3410	0.2877	0.6117	0.3490	0.3167	0.2823	0.5133
$\theta = 1.25$	0.4903	0.3090	0.3230	0.5347	0.4230	0.2667	0.2757	0.4957	0.3740	0.2323	0.2423	0.4720
$\theta = 1.5$	0.9513	0.8050	0.7987	0.9643	0.8787	0.7287	0.7353	0.9333	0.7907	0.6720	0.6800	0.9110
$n = 200$	$p = 0.1$				$p = 0.2$				$p = 0.3$			
	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}
$\theta = 0.75$	0.9337	0.7217	0.6167	0.9423	0.8933	0.6783	0.5860	0.9167	0.7307	0.6363	0.5423	0.8223
$\theta = 1.25$	0.7920	0.5513	0.5537	0.8103	0.7280	0.4823	0.4993	0.7773	0.6243	0.4277	0.4437	0.7303
$\theta = 1.5$	0.9993	0.9817	0.9750	0.9993	0.9980	0.9740	0.9623	0.9993	0.9613	0.9433	0.9233	0.9937
Model 2. Power study for $\theta = 1 + Hn^{-0.5}$												
$n = 100$	$p = 0.1$				$p = 0.2$				$p = 0.3$			
	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}
$H = -4$	0.9720	0.9960	0.9872	0.9914	0.9346	0.9928	0.9754	0.9844	0.7752	0.9822	0.9564	0.9328
$H = -2$	0.5580	0.4828	0.4746	0.6094	0.4236	0.4118	0.3966	0.5528	0.2354	0.3684	0.3510	0.4364
$H = 2$	0.4784	0.3112	0.3118	0.4954	0.4278	0.2608	0.2736	0.4730	0.3910	0.2194	0.2302	0.4578
$H = 4$	0.9594	0.8464	0.8296	0.9598	0.9160	0.7564	0.7430	0.9412	0.8204	0.6572	0.6478	0.9056
$n = 200$	$p = 0.1$				$p = 0.2$				$p = 0.3$			
	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}	T_n	$A_{(5)}$	$A_{(2)}$	T_n^{KG}
$H = -4$	0.9876	0.9938	0.9856	0.9940	0.9704	0.9820	0.9604	0.9864	0.8664	0.9682	0.9306	0.9490
$H = -2$	0.5712	0.4390	0.4406	0.5976	0.4636	0.3870	0.3808	0.5312	0.2892	0.3370	0.3338	0.4268
$H = 2$	0.5108	0.3312	0.3354	0.5284	0.4562	0.2768	0.2818	0.4962	0.3998	0.2302	0.2410	0.4736
$H = 4$	0.9686	0.8946	0.8610	0.9676	0.9446	0.8216	0.7830	0.9550	0.8764	0.7330	0.7032	0.9316

Moreover, the results on Table 2 are based on 5000 trials and show the power under alternatives for two different sample sizes $n = 100, 200$, censoring levels $p = 0.1, 0.2, 0.3$ and different values of parameter θ. In particular, for Model 2, we choose $\theta = 1 + Hn^{-0.5}$ and $H \in \{-4, -2, 2, 4\}$. Both, tables and figures, include a comparison to the Pearson-type test proposed by Akritas (1988). As before, we use the number of cells (nc) equal to 2 and 5.

The goal here is to show the changes in power when varying both θ parameter and the censoring rate p. In particular, Figure 2 is devoted to illustrate the changes in power when considering a given range for $\theta \in \Theta$. That is, Figure 2 contains the power curves of the statistic for all the alternatives, for $n = 50$ and a fixed moderate censoring rate of $p = 0.2$. On the other hand, Table 2 is devoted to show the changes in power when considering different censoring rates. Therefore, Table 2 contains the power study for the remaining sample sizes, $n = 100, 200$, for three fixed values of θ and different censoring rates $p = 0.1, 0.2, 0.3$.

Concerning the uniformity test (Model 1), it is clear that for alternatives (b) and (c) our test outperforms that proposed by Akritas. Additionally, our test neither depends on the number of cells nor on the choice of cell boundaries. The influence of the number of cells in Akritas proposal is made obvious in panels (a1)–(c3) of Figure 2. While $A_{(2)}$ gives better results than $A_{(5)}$ for alternative (a), the opposite can be observed for alternatives (b) and (c). Unfortunatelly, the modification of the maximum correlation coefficient exhibits also some weak points. That is, the alternative (a) for $\theta \in (0, 1)$ does not provide satisfactory results, since $Q = 1$ for $\theta = 0.5$. Regarding the exponentiality test (Model 2), we get better results than the competitive test of Akritas (1988) when the alternative is Weibull with parameter $\theta > 1$. For $\theta < 1$, our test reaches the high power of the Pearson-type test for big sample sizes. However, notice that in Model 2 and for all the considered values of θ, the test statistic under the Koziol-Green model, T_n^{KG}, gives very good results independently on the sample size.

6. Further extensions and application

6.1. Composite null hypothesis

So far, our test T_n has been designed to test a fully specified null hypothesis. It does strongly depend on the fact that the transformed lifetime $F_0(X)$ is $[0, 1]$ uniformly distributed under $H_0 : F = F_0$. In this section we consider a more general case, that is, when the distribution function to be tested depends on an unknown parameter λ. Let now consider the following null hypothesis

$$H_0 : F \in \{F_\lambda : \lambda \in \mathbb{R}^d\}.$$

In this case, first we need to estimate the parameter λ using, e.g., a maximum-likelihood estimator $\hat{\lambda}$. Clearly, if F_λ is twice differentiable in λ and the estimator $\hat{\lambda}$ is $\sqrt{n}$ consistent, by the Taylor expansion we have that $F_{\hat{\lambda}}(X) = U + O_P(n^{-1/2})$, where $U = F_\lambda(X) \sim U[0,1]$ under the null hypothesis H_0. The test statistic $\tilde{Q}_n$ should still admit a normal limit but the error term enters the variance of our test statistic and hence the asymptotic variance given in Theorem 1 is no longer valid. Even though the theoretical properties of our test in the case of such a composite hypothesis are beyond the scope of this paper, to test this kind of hypothesis we propose a modified test with a jackknife estimator of the variance, which does take into account the estimation of the parameters and works very well in practice. Preliminary simulation studies, as those given in Figure 3, confirm the normality of the statistic and adequacy of the variance. We proceed as follows:

1. Based on the sample $X_1, \dots, X_n$, find the maximum-likelihood estimator (MLE) $\hat{\lambda}$.
2. Define the pseudo-values $\tilde{X}_i = F_{\hat{\lambda}}(X_i)$ for $i = 1, \dots, n$.
3. Based on the sample $\tilde{X}_1, \dots, \tilde{X}_n$, compute the test statistic $\tilde{Q}_n$ defined in (7).
4. Compute the jackknife estimator of the variance following the steps:
    - For every $i = 1, \dots, n$, choose the subsample $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$ and compute the MLE $\hat{\lambda}^{(-i)}$.
    - Define the pseudo-values $\tilde{X}_j = F_{\hat{\lambda}^{(-i)}}(X_j)$ for $j = 1, \dots, i-1, i+1, \dots, n$.
    - Based on the the sample $\tilde{X}_1, \dots, \tilde{X}_{i-1}, \tilde{X}_{i+1}, \dots, \tilde{X}_n$, compute the test statistic $\tilde{Q}_n^{(-i)}$.
    - Set
    $$nV_n(\tilde{Q}_n) = (n-1)\sum_{i=1}^{n}(\tilde{Q}_n^{(-i)} - \bar{Q}_n)^2,$$
    where $\bar{Q}_n = \frac{1}{n}\sum_{i=1}^{n}\tilde{Q}_n^{(-i)}$.
5. Define the test statistic
$$J_n := \frac{\sqrt{n}\tilde{Q}_n}{\sqrt{nV_n(\tilde{Q}_n)}}.$$
6. Reject H_0 if
$$J_n \le \Phi^{-1}(\alpha/2) \text{ or } J_n \ge \Phi^{-1}(1-\alpha/2).$$

In order to check the behaviour of this new jackknife-test J_n, we study the hypothesis $H_0 : F \in \{\exp(\lambda) : \lambda \in (0,\infty)\}$, where the alternatives come from the Weibull distribution. Our simulated sample comes from $\exp(\lambda = 1)$ and λ is estimated using maximum likelihood. In Figure 3 we compare the test based in T_n, defined in equation (8)

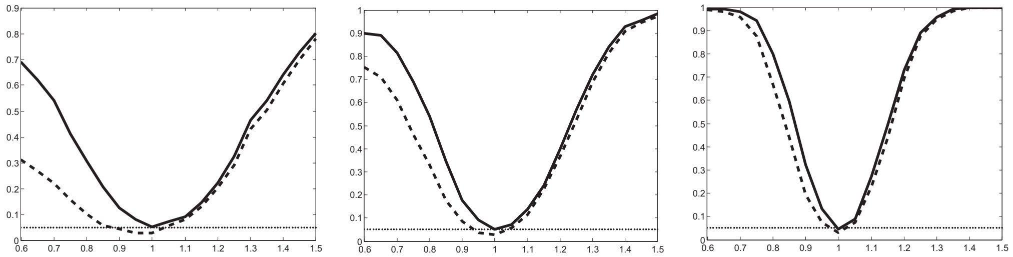


Figure 3: *Power study: J_n (solid line) and T_n (dashed line), where $n = 50$ (left), $n = 100$ (middle) and $n = 200$ (right), censoring rate $p = 0.2$*

of Section 4, with that based on J_n. As expected, the new test based on J_n gives very good results: The variance estimator adapted to the composite hypothesis is performing very well, leading to a more powerful test. The differences in power between both statistics seem to decrease with the sample size. Nevertheless, the theoretical properties of J_n are out of scope of the present paper.

6.2. Real data example

We illustrate the use of our test on the head-and-neck cancer data from Nikulin and Haghighi (2006). These authors fitted the Generalized-Power Weibull distribution $F(x,\sigma,v,\gamma)$ to the data. Motivated by the boxplot in Figure 4, we remove several observations which could be considered as outliers. This gives us 44 observations with around 11% censoring rate. We perform a goodness-of-fit test for the before-mentioned Generalized-Power Weibull distribution $F_0^a(x,\sigma,v,\gamma) = F(x,\sigma,v,\gamma)$. Additionally, we also consider the Weibull distribution $F_0^b(x,\sigma,v) = F(x,\sigma,v,1)$ and the Exponential distribution $F_0^c(x,\sigma) = F(x,\sigma,1,1)$, where

$$F(x,\sigma,v,\gamma) = 1 - \exp\left(1 - (1 + (x/\sigma)^v)^{1/\gamma}\right).$$

First, we fitted the parameters using MLE under random censoring obtaining the estimators $(\hat{\sigma},\hat{v},\hat{\gamma})$ and the following distributions $F_0^a(x,4.63,1.82,1.91)$, $F_0^b(x,1.44,8.45)$ and $F_0^c(x,8.33)$. Then we applied our test J_n and obtained the following p-values: $p^a = 0.86$, $p^b = 0.88$ and $p^c = 0.01$ for the Generalized-Power Weibull, Weibull and Exponential, respectively. Hence, the results of the test confirm what Figure 4 shows, that both Generalized-Power Weibull and Weibull fit the data very well, whereas the Exponential distribution is not adequate to describe the head-and-neck cancer data.

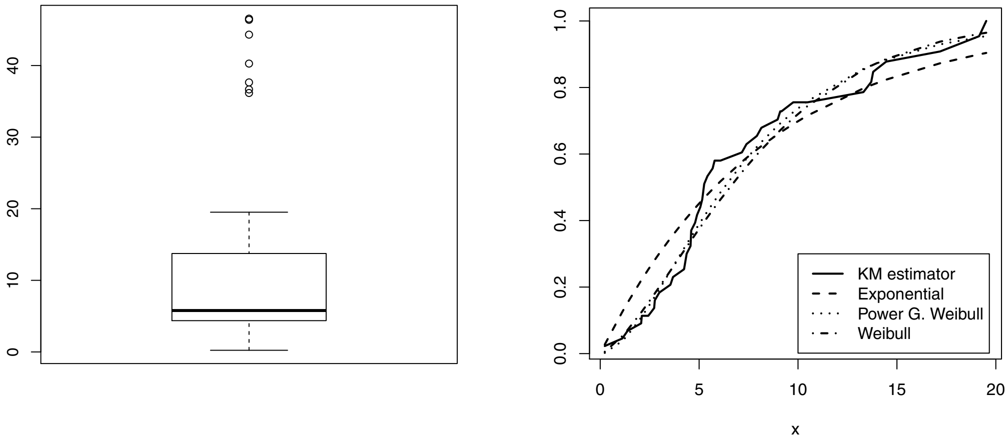


Figure 4: Boxplot (left) and Kaplan-Meier estimator (right) together with $F_0^a(x,\hat{\sigma},\hat{\nu},\hat{\gamma})$ (dashed), $F_0^b(x,\hat{\sigma},\hat{\nu})$ (dotted) and $F_0^c(x,\hat{\sigma})$ (dot-dashed) for the head-and-neck cancer data.

7. Conclusions

In this work we developed and studied a goodness-of-fit test based on maximum correlation under random censoring. The advantage of our test over other goodness-of-fit competitors, like χ^2 test proposed by Akritas (1988), is its simplicity. Our test is asymptotically normally distributed and neither the number of classes nor the class boundaries have to be chosen. The proposed test outperforms that by Akritas (1988) for most of the alternatives studied. Even though the test was designed to check uniformity, with a simple transformation it can be applied to any, fully specified, continuous distribution. Finally, it can be extended to composite hypothesis, that is, when the distribution in the null hypothesis is known up to a parameter. A jackknife modification for the asymptotic variance has been proposed. A theoretical study of the test under the composite null hypothesis is out of the scope of the present paper and purpose of further research.

8. Appendix

Proof of Theorem 1

In view of (7), we can write $\tilde{Q}_n$ in the following way

$$\tilde{Q}_n = \int_0^1 \int_0^1 \tilde{h}(x,y) F_n(dx) F_n(dy), \tag{10}$$

where

$$\tilde{h}(x,y) = (6x-2)1_{\{y<x\}} - 6x1_{\{y>x\}}.$$

In the fist step of the proof we write $\tilde{Q}_n$ as a sum of four terms as follows

$$\tilde{Q}_n \quad = \quad \tilde{Q}_1 + \tilde{Q}_{2n} + \tilde{Q}_{3n} + \tilde{Q}_{4n},$$

where

$$\begin{aligned}
\tilde{Q}_1 &= \int_0^1 \int_0^1 \tilde{h}(x,y)F(dx)F(dy) \\
\tilde{Q}_{2n} &= \int_0^1 \int_0^1 \tilde{h}(x,y)(F_n(dx)-F(dx))F(dy) \\
\tilde{Q}_{3n} &= \int_0^1 \int_0^1 \tilde{h}(x,y)(F_n(dy)-F(dy))F(dx) \\
\tilde{Q}_{4n} &= \int_0^1 \int_0^1 \tilde{h}(x,y)(F_n(dx)-F(dx))(F_n(dy)-F(dy)).
\end{aligned}$$

By (6) and since F is continuous, we have that $\tilde{Q}_1 = Q^1$. As to $\tilde{Q}_{2n} + \tilde{Q}_{3n}$, we obtain

$$\tilde{Q}_{2n} + \tilde{Q}_{3n} = \int_0^1 \varphi(x)(F_n(dx) - F(dx)),$$

where

$$\begin{aligned}
\varphi(x) &= \int_0^1 \tilde{h}(y,x)F(dy) + \int_0^1 \tilde{h}(x,y)F(dy) \\
&= 12xF(x) - 6x - 2 - 12\int_0^x yF(dy) + 6\int_0^1 yF(dy).
\end{aligned}$$

It remains to show that $\tilde{Q}_{4n} = o_{\mathbb{P}}(n^{1/2})$. For this, set $\tau_{\tilde{H}} = \inf\{t : \tilde{H}(t) = 1\}$, where $\tilde{H}(t) = \mathbb{P}(X \leq t)$ is the cdf of the observed sample. Since the support $\text{supp}(F) \in [0,1]$ and G fulfills assumption A1, we have that $\tau_{\tilde{H}} = 1$. Moreover, by definition of $\tilde{h}(x,y)$, we can show that

$$\tilde{Q}_{4n} = -12\int_0^1 x(F_n(x)-F(x))(F_n(dx)-F(dx)) - 2(F_n(1)-F(1))^2 =: \tilde{Q}_{4n}^a + \tilde{Q}_{4n}^b.$$

Now, we may consider the two terms, $\tilde{Q}_{4n}^a$ and $\tilde{Q}_{4n}^b$, separately. According to Theorem 2 (7) in Ying (1989) and under A1, the process $\sqrt{n}(F_n - F)$ converges weakly to a Brownian process. See, also equation (11) in Wellner (2007). More precisely,

$$\sqrt{n}(F_n - F) \to (1-F)\mathbb{B}(C), \text{ in } D[0,\tau_{\tilde{H}}],$$

where $\mathbb{B}(C)$ is a Brownian process and $D[0,\tau_{\tilde{H}}]$ denotes the Skorohod space. Furthermore, since F is continuous and D^0 is a set of uniformly bounded functions, we have that $\sqrt{n}(F_n - F) \in D^0$ with probability exceeding $1-\varepsilon$ for every $\varepsilon > 0$. Additionally, $x \in [0,1]$ and $\sup_{x\in[0,\tau_{\tilde{H}}]} |F_n(x) - F(x)| \to 0$ almost surely. Hence, using Theorem 2.1 in Rao (1962) with $g(x) = \sqrt{n}(F_n(x) - F(x))x$, we obtain

$$\sqrt{n}\tilde{Q}^a_{4n} = -12\int_0^1 g(x)(F_n(dx) - F(dx)) = o_{\mathbb{P}}(1).$$

Additionally, under A1, $F_n(1) - F(1) = O_P(\sqrt{n})$ and hence $\sqrt{n}\tilde{Q}^b_{4n} = o_{\mathbb{P}}(1)$. Notice that, $F_n(1) - F(1) = \int_0^1 1(F_n(dx) - F(dx))$. Hence we apply the results from Stute (1995) for $\varphi(x) = 1$.

Finally, we obtain the following representation

$$\tilde{Q}_n = Q^1 + \int_0^1 \varphi(x)(F_n(dx) - F(dx)) + o_{\mathbb{P}}(n^{1/2}).$$

The asymptotic normality is now a direct consequence of Stute (1995). More precisely, under A1 and A2, we obtain

$$\sqrt{n}\int_0^1 \varphi(x)(F_n(dx) - F(dx)) \to \mathcal{N}(0,\sigma^2).$$

This completes the proof. ■

Proof of Lemma 1

Recall, that

$$\begin{aligned}\sigma_n^2 = &\int_0^1 \frac{\varphi_n^2(x)}{1-G_n(x-)}F_n(dx) - \left[\int_0^1 \varphi_n(x)F_n(dx)\right]^2 - \\ &-\int_0^1 \left[\int_x^1 \varphi_n(y)F_n(dy)\right]^2 \frac{(1-F_n(x))G_n(dx)}{(1-H_n(x))^2} \\ &=: A_{1n} - A_{2n} - A_{3n}.\end{aligned}$$

where

$$\varphi_n(x) = 12xF_n(x) - 6x - 2 - 12\int_0^x yF_n(dy) + 6\int_0^1 yF_n(dy).$$

By consistency of the Kaplan-Meier estimator, we have $\varphi_n(x) \to \varphi(x)$ in probability. Let us consider the first term in the representation of σ_n^2. Since $|\varphi_n(x)| \leq K_1 =$ constant, $G_n(x) \to G(x)$ in probability,

$$\max_{i=1,\ldots,n} \frac{1-G(X_i-)}{1-G_n(X_i-)} = O_P(1)$$

by Zhou (1991) and

$$\frac{1}{1-G_n(x-)} = \frac{1}{1-G(x-)} + \frac{G_n(x-)-G(x-)}{(1-G_n(x-))(1-G(x-))}$$

we have

$$A_{1n} = \int_0^1 \frac{\varphi^2(x)}{1-G(x)} F_n(dx) + o_P(1)$$

Finally, by Theorem 1.1. in Stute and Wang (1993), $\int \Phi(x)F_n(dx) \to \int \Phi(x)F(dx)$ with probability 1 and hence in probability, provided that $\int |\Phi(x)|F(dx) < \infty$. Hence, by A3, we obtain

$$A_{1n} \xrightarrow{P} \int_0^1 \frac{\varphi^2(x)}{1-G(x)} F(dx).$$

Obviously, we have

$$A_{2n} \xrightarrow{P} \left[\int_0^1 \varphi(x)F(dx)\right]^2.$$

Finally, similarly as we have done for A_{1n}, we may show that

$$A_{3n} = \int_0^1 \left[\int_x^1 \varphi(y)F(dy)\right]^2 \frac{(1-F(x))G_n(dx)}{(1-H(x))^2} + o_P(1)$$

By A3 and since $|\varphi(y)| \leq K_2 =$ constant we obtain

$$\int_0^1 \left[\int_x^1 \varphi(y)F(dy)\right]^2 \frac{(1-F(x))G(dx)}{(1-H(x))^2} \leq K_2^2$$
$$\int_0^1 \left[\int_x^1 \frac{F(dy)}{(1-G(y))^{1+\varepsilon}}\right] \frac{G(dx)}{(1-G(x))^{1-\varepsilon}} < \infty.$$

Hence, by Theorem 1.1. in Stute and Wang (1993),

$$A_{3n} \to \int_0^1 \left[\int_x^1 \varphi(y)F(dy)\right]^2 \frac{(1-F(x))G(dx)}{(1-H(x))^2}$$

in probability. This completes the proof. ■

Acknowledgements

This work has been supported by Spanish grant MTM2014-56535-R (Spanish Ministry of Economy and Competitiveness).

References

Akritas, M.G. (1988). Pearson-type goodness-of-fit test: the univariate case. *Journal of the American Statistical Association*, 83, 222–230.

Balakrishnan, N., Chimitova E. and Vedernikova, M. (2015). An empirical analysis of some nonparametric goodness-of-fit tests for censored data. *Communications in Statistics-Simulation and Computation*, 44, 1101–1115.

Cuadras, C.M. and Fortiana, J. (1993). Continuous metric scaling and prediction. *Multivariate Analysis, Future Directions*, vol. 2 (eds. C.M. Cuadras and C.R. Rao), 47–66. Amsterdam: North-Holland.

Darling, D.A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises tests. *The Annals of Mathematical Statistics*, 28, 823–838.

Fortiana, J. and Grané, A. (2003). Goodness-of-fit tests based on maximum correlations and their orthogonal decompositions. *Journal of the Royal Statistical Society: Series B*, 65, 115–126.

Grané, A. (2012). Exact goodness-of-fit tests for censored data. *Annals of the Institute of Statistical Mathematics*, 64, 1187–1203.

Grané, A. and Tchirina, A. (2013). Asymptotic properties of a goodness-of-fit test based on maximum correlations. *Statistics*, 47, 202–215.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.

Koziol, J.A. and Green S.B. (1976). A Cramer-von Mises statistic for randomly censored data. *Biometrika*, 63, 465–474.

Massey, Jr, F.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46, 68–78.

Nikulin, M. and Haghighi, F. (2006). A chi-squared test for the generalized power Weibull family for the head-and-neck cancer censored data. *Journal of Mathematical Sciences*, 133, 1333–1341.

Novoa-Muñoz, F. and Jiménez-Gamero, M.D. (2016). A goodness-of-fit test for the multivariate Poisson distribution. *SORT*, 40, 1–26.

Rao, R.R. (1962). Relations between weak and uniform convergence of measures with applications. *Annals of Mathematical Statistics*, 33, 659–680.

Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *Annals of Statistics*, 21, 1591–1607.

Stute, W. (1994). The bias of Kaplan-Meier integrals. *Scandinavian Journal of Statistics*, 21, 475–484.

Stute, W. (1995). The central limit theorem under random censorship. *Annals of Statistics*, 23, 422–439.

Torabi, H., Montazeri, N. H. and Grané, A. (2016). A test for normality based on the empirical distribution function. *SORT*, 40, 55–88.

Wellner, J.A. (2007). On an exponential bound for the Kaplan-Meier estimator. *Lifetime Data Analysis*, 13, 481–496.

Ying, Z. (1989). A note on the asymptotic properties of the product-limit estimator on the whole line. *Statistics & Probability Letters*, 7, 311–314.

Zhou, M. (1991). Some properties of the Kaplan-Meier estimator for independent nonidentically distributed random variables. *The Annals of Statistics*, 19, 2266–2274.

SORT 41 (1) January-June 2017, 159-176 DOI: 10.2436/20.8080.02.56

A Bayesian stochastic SIRS model with a vaccination strategy for the analysis of respiratory syncytial virus

M. Jornet-Sanz,[1] A. Corberán-Vallet,[1] F.J. Santonja[1,*] and R.J. Villanueva[2]

Abstract

Our objective in this paper is to model the dynamics of respiratory syncytial virus in the region of Valencia (Spain) and analyse the effect of vaccination strategies from a health-economic point of view. Compartmental mathematical models based on differential equations are commonly used in epidemiology to both understand the underlying mechanisms that influence disease transmission and analyse the impact of vaccination programs. However, a recently proposed Bayesian stochastic susceptible-infected-recovered-susceptible model in discrete-time provided an improved and more natural description of disease dynamics. In this work, we propose an extension of that stochastic model that allows us to simulate and assess the effect of a vaccination strategy that consists on vaccinating a proportion of newborns.

MSC: 62P10.

Keywords: Infectious diseases, respiratory syncytial virus (RSV), discrete-time epidemic model, stochastic compartmental model, Bayesian analysis, intervention strategies.

1. Introduction

Effective surveillance and control measures are essential to protect public health by rapidly detecting and responding to outbreaks of infectious diseases, which pose a growing threat to human health. Shortcomings in surveillance, vaccines and treatment can result in rising morbidity and mortality. Innovative surveillance methods have been

[*] e-mail: Francisco.Santonja@uv.es

[1] Department of Statistics and Operational Research. Universitat de València. Faculty of Mathematics. Dr. Moliner, 50. 46100 Burjassot, Spain.

[2] Institute for Multidisciplinary Mathematics. Universitat Politècnica de València. Building 8G, 2 Floor, Camino de Vera s/n. 46022 Valencia, Spain.

Received: July 2016
Accepted: January 2017

recently developed in an effort to improve outbreak detection capabilities. Numerous epidemiological models have also been proposed to simulate and analyse the impact of different vaccination strategies from an economic and sanitary point of view. Nowadays, the use of models is considered as an effective tool to both represent the evolution of diseases and assess the impact of control interventions (World Health Organization, 2016).

Most of the approaches that are currently used to study the impact of vaccination programs fall into one of the following two categories: compartmental mathematical models (Acedo et al., 2010; Hogan et al., 2016; Van Hoek et al., 2011; Christensen et al., 2013; Yu et al., 2016) or computer models defined through complex schemes of interaction (Pérez-Breva et al., 2014; Vannia et al., 2012; Craig et al., 2014; Poletti et al., 2015). Compartmental models divide the population being studied into a set of distinct compartments according to the disease status (for instance, the susceptible-infected-recovered model divides the population into three categories) and model the evolution of infectious diseases through changes in the number of individuals in each compartment. They are usually based on ordinary differential equations, which imply a continuous-time deterministic model. Besides, they are defined assuming that all the individuals in the population are equally likely to contact any other individual (Ma and Li, 2009; Brauer, 2008). However, contact patterns in real populations are indeed more heterogeneous. Therefore, models involving homogeneous mixing should be replaced by models incorporating stochastic effects (Brauer, 2008). Stochastic models are able to accommodate the stochasticity inherent in the transmission of infection by considering that the number of individuals in each compartment is a random variable with its associated probability distribution (Allen, 2008). In addition, stochastic models can be easily analysed from a Bayesian viewpoint (see, for example, Gibson and Renshaw, 1998; O'Neill, 2002; Boys and Giles, 2007; Weidemann et al., 2014).

A Bayesian stochastic susceptible-infected-recovered-susceptible (SIRS) model in discrete time has been recently proposed to model respiratory syncytial virus (RSV) dynamics in the region of Valencia, Spain (Corberán-Vallet and Santonja, 2014). The proposed model, which can be seen as a discrete time Markov chain model (Allen, 2008), does not imply mass-action mixing of individuals in the population. In addition, the probability of disease transmission depends on a transmission rate that is allowed to vary stochastically over time. This feature is fundamental to provide an accurate representation of the disease dynamics.

RSV is the most important cause of lower respiratory tract illness in infants and children worldwide. It causes repeat infections throughout life and significant disease in pediatric and elderly population. Due to the high burden of disease globally, RSV has been a priority for vaccine development. However, efforts to develop a safe and effective vaccine have yet to lead to a licensed product (Anderson et al., 2013; Jones et al., 2014; Higgings, Trujillo and Keech, 2016; Roberts et al., 2016). The epidemiology and burden of RSV disease point to several target populations for vaccines, which may require different vaccination strategies according to the age. The highest priority target population are infants < 6 months of age who are at highest risk of severe disease. The enhanced disease observed after a formalin-inactivated RSV (FI-RSV) vaccine directed development of RSV vaccines toward live virus vaccines. Yet these young infants present challenges to vaccine development. They may not respond well to a vaccine because of immature immune system, suppression of the immune response by presence of maternal antibody, and an elevated susceptibility to disease with live RSV infection. The second target population are children ≥ 6 months of age, both to prevent their disease and potential transmission to younger children. The third target population are pregnant women, since high titers of maternally derived RSV antibody have been shown to correlate inversely with the incidence and severity of RSV infection in the first six months of life. This maternal vaccination strategy would protect newborns both by placental transfer of antibodies and by blocking transmission (Dudas and Karron, 1998). However, it would not provide protection for children beyond 4–6 months of age, and so this strategy would be followed by direct child vaccination as maternal antibody wanes. The last target population are the elderly, who are also at risk for severe disease. See Higgings et al., 2016 for a current summary of RSV vaccine research and development.

Taking into account recommendations for RSV vaccine development (Anderson et al., 2013), we present here an extension of the model proposed in (Corberán-Vallet and Santonja, 2014) that allows us to simulate and assess the impact of vaccination programs. Because most efforts are directed towards immunizing infants from birth to six months (Beeler and Eichelberger, 2013), the simulated strategy consists on vaccinating a proportion of newborns. This strategy is similar to the one implemented in (Acedo et al., 2010).

2. Case study

Our study focuses on weekly hospitalizations for RSV-related illnesses among children younger than two years of age in the Spanish region of Valencia. Children aged < 2 years are the target population most problematic by possible severe complications. On some occasions, hospitalization may be necessary, especially for RSV bronchiolitis and pneumonia.

In particular, we have data on the number of new hospitalizations per week from week beginning January 1st 2001 to week beginning December 20th 2004 (see Figure 1). As can be seen, RSV activity presents a clear seasonal pattern: epidemics occur yearly between late fall and early spring.

Acedo et al. (Acedo et al., 2010) proposed a compartmental model based on ordinary differential equations to describe these data and perform a vaccination program analysis. They assumed that the sizes of the susceptible, infected, removed and vaccinated populations were large enough so that the mixing of individuals in the population was homogeneous. However, that is not the case, specially at the beginning of epidemics. In

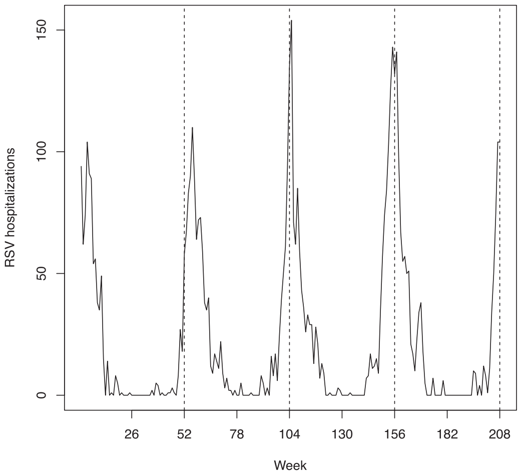


***Figure 1:** Hospitalizations for respiratory syncytial virus (RSV) infection among children aged* < 2 *years from week beginning January* $1st$ 2001 *to week beginning December* $20th$ 2004 *in the Spanish region of Valencia.*

addition, the transmission rate was modeled assuming the same seasonal pattern for all the years, which is not a realistic description of the data.

Corberán-Vallet and Santonja (Corberán-Vallet and Santonja, 2014) proposed a stochastic SIRS model in discrete time that provided a precise representation of the pattern of disease. That model was also able to quite accurately identify the start of a new RSV epidemic and its increase. However, vaccination strategies were not studied. Similar to that study, we also confine our analysis to data collected from week beginning July 2nd, 2001 (week 27 in the time plot). Since no child was hospitalized the week before, we can assume that the susceptible population at this time period is the population of children aged < 2 in the region of Valencia.

It is important to emphasize here that the interest when simulating the effect of vaccination strategies may be to study the decrease in disease incidence. In this case study, the available data refer to hospitalizations. Let i_t and y_t be, respectively, the number of infections and hospitalizations at time t. It is reasonably to assume the following relationship:

$$y_t \sim Bin(i_t, \rho)$$

where ρ is the probability of being hospitalized for RSV infection. Because information regarding the number of newly infected children per week is not available, it is not possible to make a statistical robust estimate of ρ. In Spain, the percentage of children who require hospitalization for RSV is around 0.5% and 2% of the number of infected children (Contreras, 2016; Parra et al., 2013). This percentage coincides with the results obtained in Acedo et al. (2010). Hence, if the interest relies on analysing the number of infections, we can assume that the number of infected children at week t ($t = 1, 2, \ldots, T = 208$) is given by the number of hospitalized children divided by the hospitalization rate:

$$i_t = y_t / h$$

with $h = 0.02$ as proposed in Acedo et al. (2010).

3. Model formulation

3.1. Model formulation without vaccination

In this section we describe a Bayesian stochastic susceptible-infected-recovered-susceptible (SIRS) model in discrete time that was proposed by Corberán-Vallet and Santonja (2014) to study infectious disease dynamics. Let i_t denote the number of infected children at week t, $t = 1, 2, \dots, T$. Because the population of children aged < 2 years in the region of Valencia is finite, the observations are assumed to be Binomial distributed:

$$i_t \sim Bin(S_{t-1}, p_t) \tag{1}$$

where S_{t-1} represents the susceptible population at time $t-1$; that is, the number of individuals not yet infected with the disease at time $t-1$; and p_t is the probability of becoming infected at time t.

In this discrete-time model, the number of individuals in each compartment is examined at discrete time steps. Using a fixed population, the number of susceptible (S), infected (I), and recovered (R) individuals at time t are updated through the following recursive equations:

$$\begin{aligned} S_t &= S_{t-1} - i_t + bR_{t-1} + n_t - \frac{S_{t-1}}{N} n_t \\ I_t &= I_{t-1} - aI_{t-1} + i_t - \frac{I_{t-1}}{N} n_t \\ R_t &= R_{t-1} - bR_{t-1} + aI_{t-1} - \frac{R_{t-1}}{N} n_t \end{aligned} \tag{2}$$

where a is the proportion of infected individuals that recover per unit time; b is the proportion of recovered individuals who lose their immunity and become susceptible again per unit time; n_t is the number of births at time t; and N is the constant population size. Taking into account that the average time to recover from RSV illness is 10 days and the average time to lose immunity is 200 days, we can set $a = \frac{7 \text{ days (one week)}}{10 \text{ days (recover time)}} = 0.7$ and $b = \frac{7 \text{ days (one week)}}{200 \text{ days (time to lose immunity)}} = 0.035$ (Acedo et al., 2010). Because the recovery time and time to lose immunity for RSV are well-known, we have considered these values as deterministic inputs. Otherwise, these quantities should be considered as additional parameters of the model with their corresponding prior distribution. In addition, using demographic data from the Spanish National Institute of Statistics (http://www.ine.es), the average weekly number of births in the region of Valencia for years 2001–2004 and the population size can be set $n_t = 879$ and $N = 83,205$. The flowchart diagram for the model is described in Figure 2.

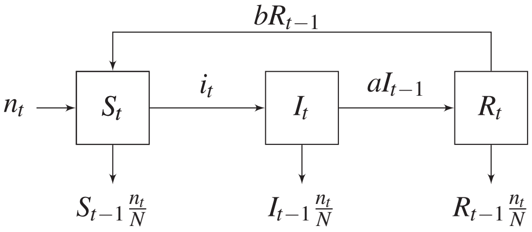


Figure 2: *Flowchart of the SIRS model without vaccination. Boxes represent compartments and arrows represent transitions between compartments, labelled by the parameters of the model.*

The probability p_t was modelled as:

$$p_t = \min\left\{\frac{i_{t-1}^{\alpha}\exp\{r_t\}}{1+i_{t-1}^{\alpha}\exp\{r_t\}} + c, 1\right\} \tag{3}$$

where the mixing parameter α allows for heterogeneous mixing (homogeneous mixing corresponds to $\alpha = 1$ (Bjørnstad, Finkenstädt and Grenfell, 2002); c represents a constant probability of becoming infected and so it accounts for the occurrence of new cases after the disease has faded out; and $\exp\{r_t\}$ represents the time-varying transmission rate. To accommodate the seasonal pattern observed in the dynamics of RSV, this transmission rate is modelled by means of sine-cosine waves as:

$$r_t = a_0 + \sum_{k=1}^{K}\left[a_{2k-1}\sin\left(\frac{2k\pi t}{52}\right) + a_{2k}\cos\left(\frac{2k\pi t}{52}\right)\right] + \epsilon_t \tag{4}$$

where ϵ_t is a random effect that represents unspecified features of week t. Note that this formulation ensures that the probability lies in the interval 0–1 and it also takes into account the transmissible nature of the infection. The value of K depends on the data under study and it is set as the highest value k^* so that the corresponding parameters a_{2k^*-1} and a_{2k^*} are significant.

The parameters of the model are α, c, $\{a_k\}_{k=0}^{2K}$, and $\{\epsilon_t\}_{t=1}^{T}$. The prior distribution assumed for parameter α is the Uniform distribution in the interval 0–1. The Uniform distribution in the interval 0–0.01 is assigned to parameter c. In this case study, this range of variation for c is enough to capture the probability of infection the first week of epidemic periods. However, a wider range may be necessary in the analysis of different diseases. Parameters $\{a_k\}$ are assumed to have zero mean Gaussian distributions with standard deviations σ_{a_k}; and $\{\epsilon_t\}$ are Gaussian distributed random effects with zero mean and standard deviations σ_ϵ. All the standard deviations in the previous equations are assigned the Uniform distribution in the interval 0–5 (Gelman, 2006).

3.2. Model formulation with vaccination

We propose here an extension of the model previously described to accommodate a vaccination strategy that consists on vaccinating a proportion of newborns. As mentioned in the introduction section, the planning of effective vaccine strategies to protect infants from birth to six mounth are needed. Let $\tilde{i}_t$ be the number of infected children at time t after implementation of the vaccination program for infants. It is important to emphasize that in this section we are working with a hypothetical scenario (since there is not a RSV vaccination strategy implemented in the Community of Valencia), and so data corresponding to the number of new infections are not available. Let $\tilde{S}_{t-1}$ represent the susceptible population at time $t-1$ and $\tilde{p}_t$ the new probability of becoming infected at time t after introducing the RSV vaccine. We can assume then that:

$$\tilde{i}_t = \tilde{S}_{t-1}\tilde{p}_t \tag{5}$$

The number of individuals in each compartment is updated through the following recursive equations:

$$\begin{aligned}
\tilde{S}_t &= \tilde{S}_{t-1} - \tilde{i}_t + b\tilde{R}_{t-1} + (1-\nu)n_t - \frac{\tilde{S}_{t-1}}{N}n_t \\
\tilde{I}_t &= \tilde{I}_{t-1} - a\tilde{I}_{t-1} + \tilde{i}_t - \frac{\tilde{I}_{t-1}}{N}n_t \\
\tilde{R}_t &= \tilde{R}_{t-1} - b\tilde{R}_{t-1} + a\tilde{I}_{t-1} - \frac{\tilde{R}_{t-1}}{N}n_t \\
\tilde{V}_t &= \tilde{V}_{t-1} + \nu n_t - \frac{\tilde{V}_{t-1}}{N}n_t
\end{aligned} \tag{6}$$

Similar to Equations (2), a is the proportion of infected individuals that recover per unit time; b is the proportion of recovered individuals who lose their immunity per unit time; n_t is the number of births at time t; and N is the constant population size. Parameter ν represents the proportion of newborns that are vaccinated. We assume here that infants receive additional booster doses if necessary to induce optimal levels of RSV neutralizing antibody, and so vaccinated children do not evolve to the susceptible population. Based on this assumption, there is not transition between the vaccinated subpopulation (V) and the susceptible one (S). The flowchart diagram for the model with vaccination is described in Figure 3.

To estimate $\tilde{i}_t$ we need to know the value of $\tilde{S}_{t-1}$ and $\tilde{p}_t$. The first term is derived by applying the previous recursive equations and the probability $\tilde{p}_t$ can be estimated using the expression:

$$\tilde{p}_t = \min\left\{ \frac{\tilde{i}_{t-1}^{\alpha^*}\exp\{r_t^*\}}{1+\tilde{i}_{t-1}^{\alpha^*}\exp\{r_t^*\}} + c^*, 1 \right\} \tag{7}$$

where α^*, c^*, and $\{r_t^*\}$ represent the posterior mean estimates obtained when the model without vaccination is fitted to real data. Note that these parameters represent features of RSV dynamics that do not depend on the number of infected children, and so it is sensible to use these estimates to calculate $\tilde{p}_t$. Varying the value of ν, it is possible to find out the effect of this vaccination strategy on the number of RSV infections.

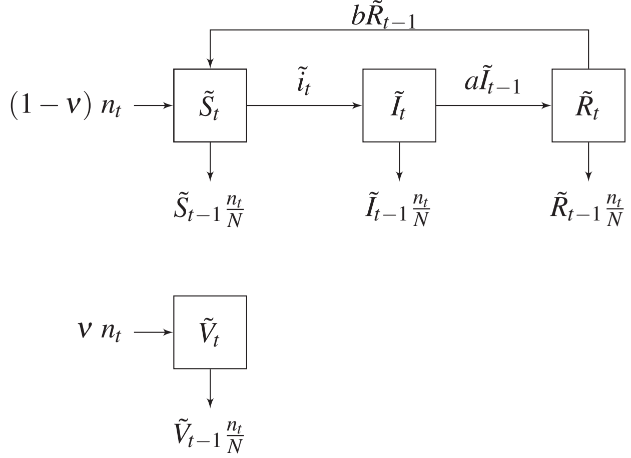


Figure 3: *Flowchart of the SIRS model with vaccination (vaccinated children do not evolve to the susceptible population). Boxes represent compartments and arrows represent transitions between compartments, labelled by the parameters of the model.*

If booster doses are not planned, vaccinated children may evolve to the susceptible population after an immunization period. In that case, the recursive equations would be replaced by:

$$\begin{aligned}
\tilde{S}_t &= \tilde{S}_{t-1} - \tilde{i}_t + b\tilde{R}_{t-1} + b^*\tilde{V}_{t-1} + (1-\nu)n_t - \frac{\tilde{S}_{t-1}}{N}n_t \\
\tilde{I}_t &= \tilde{I}_{t-1} - a\tilde{I}_{t-1} + \tilde{i}_t - \frac{\tilde{I}_{t-1}}{N}n_t \\
\tilde{R}_t &= \tilde{R}_{t-1} - b\tilde{R}_{t-1} + a\tilde{I}_{t-1} - \frac{\tilde{R}_{t-1}}{N}n_t \\
\tilde{V}_t &= \tilde{V}_{t-1} - b^*\tilde{V}_{t-1} + \nu n_t - \frac{\tilde{V}_{t-1}}{N}n_t
\end{aligned} \tag{8}$$

where b^* represents the proportion of vaccinated children who lose their immunity and become susceptible per unit time. In Acedo et al. (2010), the authors assumed an immunization period by vaccination equal to the immunization after infection. Taking into account this consideration, a value of b^* equal to 0.035 could be assumed. The flowchart diagram for this new scenario is presented in Figure 4.

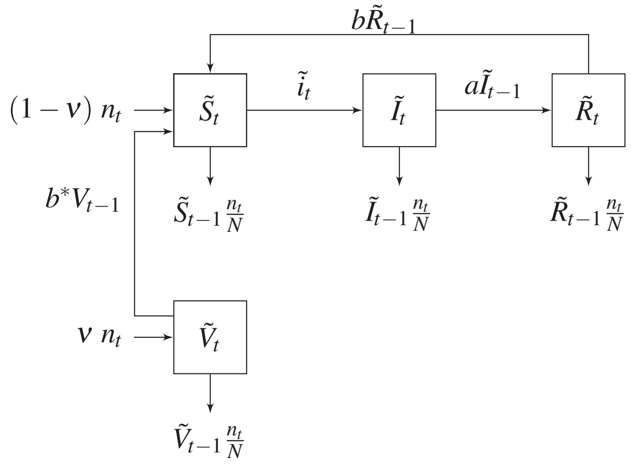


Figure 4: *Flowchart of the SIRS model with vaccination (vaccinated children may evolve to the susceptible population after an immunization period). Boxes represent compartments and arrows represent transitions between compartments, labelled by the parameters of the model.*

It is important to mention that the previously proposed model is also valid for a maternal vaccination strategy. In that case, ν would represent the proportion of newborns whose mothers have been vaccinated and so they are protected from RSV. Since this maternal immunization strategy does not provide complete protection, children in the V compartment will evolve to the susceptible population after 4–6 months (Higgings et al., 2016). The recursive equations given by (8) should then be used, with parameter b^* adapted to this immunization period.

4. Results

In this section we describe the main results obtained in the analysis of RSV data. We first show the results when the model without vaccines is fitted to the data. The Bayesian analysis of the model allows us to calculate the posterior distribution of the model parameters. Because this posterior distribution is not analytically tractable, we approached it by simulation. In particular, we obtained a random sample from it using Markov chain Monte Carlo (MCMC) simulation techniques as implemented in the free statistical software WinBUGS (Lunn et al., 2000). We fixed a burn-in period of 150000 iterations to assess the convergence of MCMC chains. To reduce the correlation for the samples, we kept one posterior sample in 25 iterations after the burn-in period until a set of 5000 iterations was obtained.

Similar to the study in Corberán-Vallet and Santonja (2014), we model directly the weekly number of RSV hospitalizations as $y_t \sim Bi(S_{t-1}, p_t)$, and so p_t represents the probability of being hospitalized at time t. The posterior mean and 95% credible intervals are displayed in Figure 5.

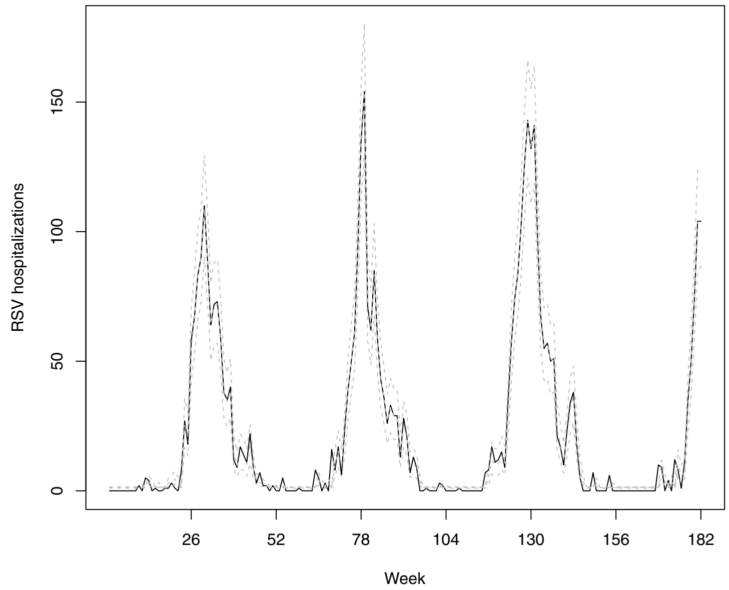


Figure 5: Hospitalizations for RSV (solid black line) together with posterior means and 95% credible intervals (dotted gray lines) from week beginning July 2nd 2001 to week beginning December 20th 2004 in the Spanish region of Valencia.

Figure 6 shows the estimated transmission rate $\exp\{r_t\}$ together with its seasonal component, which is defined by the sum of two harmonic waves ($K = 2$; higher-order frequencies were no significant). As can be seen, even though seasonality plays an important role in disease transmission, adding random effects in the transmission rate model to account for overdispersion is fundamental to provide a more realistic description of the transmission pattern.

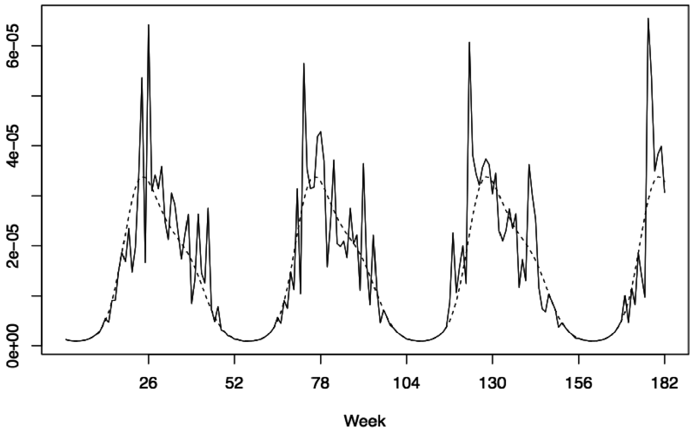


Figure 6: Estimated transmission rate together with its seasonal component from week beginning July 2nd 2001 to week beginning December 20th 2004.

Table 1: *Posterior means and 95% credible intervals for the parameters of the model.*

Parameter	Mean	Credible interval
α	0.80	(0.66,0.95)
c	1.54e-05	(1.26e-05,1.84e-05)
a_0	-11.73	(-12.28,-11.26)
a_1	-0.61	(-0.94,-0.30)
a_2	-1.55	(-2.11,-1.06)
a_3	-0.44	(-0.73,-0.17)
a_4	-0.14	(-0.42,0.10)

It is important to mention that these results are very similar to the ones presented in Corberán-Vallet and Santonja (2014). The main difference is that here we are not interested in prediction, and so we do not keep the last weeks to measure the out-of-sample forecast accuracy. Because the data set is not exactly the same, some minor differences are observed in the posterior distribution of the model parameters. The posterior mean and 95% credible intervals for the parameters of the model are shown in Table 1.

Once the posterior means of the parameters of the model without vaccines have been estimated, we can analyse the effect of the newborn vaccination strategy. As explained in Section 3.2, parameters α, c, and $\{r_t\}$ represent features of RSV dynamics that do not depend on the number of infected children, and they are used to compute both the new probability of hospitalization once the vaccine has been implemented and the new number of infections. In our simulation of the vaccine implementation, we assume that there were no vaccinations before July 2nd 2001; that is, vaccines are introduced the first week of our time frame and so we set $V_0 = 0$.

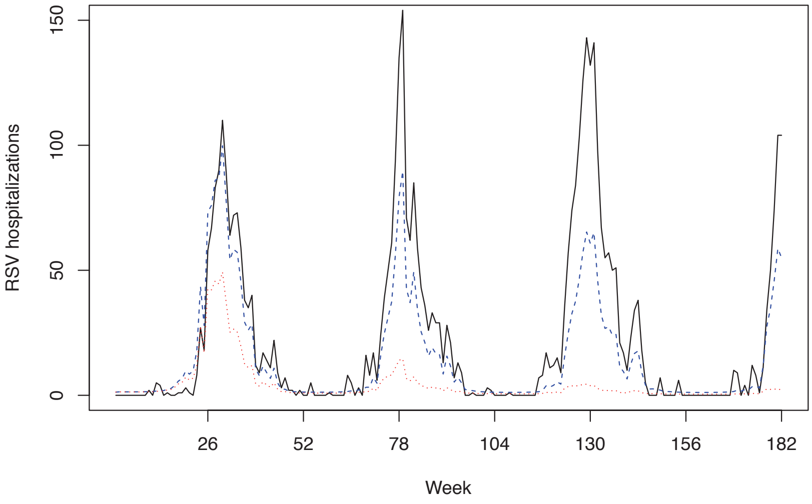


Figure 7: *Number of real hospitalizations for RSV (solid line) and simulated numbers of hospitalizations for two different coverage rates (percentages of vaccinated newborns), $\nu = 0.2$ (dashed line) and $\nu = 0.8$ (dotted line).*

Figure 7 shows the real number of hospitalizations from week beginning July 2nd 2001 to week beginning December 20th 2004 and the simulated numbers for two different values of ν: $\nu = 0.2$ and $\nu = 0.8$. As expected, the number of RSV hospitalizations decreases as the percentage of vaccinated newborns increases. Note that the value of ν is decided by policymakers and we just set these values as an example.

Taking into account an average of 6.28 hospitalization days for every infected child and €500 per day and child hospitalized (Acedo et al., 2010), we can estimate the total cost of hospitalizations for the time period of study. If no child is vaccinated, the cost of hospitalizations is approximately €$13,213,120$. This quantity decreases as ν increases. The hospitalization cost for a value of ν equal to 0.2 would be around 8.5 millions of euros, and if we set $\nu = 0.8$, 2.5 millions of euros. Note that in order to complete the economic analysis, we should also consider the vaccine price. For instance, (Acedo et al., 2010) assumed a cost of €300 per child.

4.1. Comparison with a deterministic continuous-time model

In Corberán-Vallet and Santonja (2014), the authors compared the model described in Section 3.1 with four alternative formulations of the SIRS model in discrete time: a stochastic model with a deterministic seasonal transmission rate, a stochastic model where the transmission rate was assumed to be constant over time, and the equivalent deterministic formulations. The results showed that the proposed Bayesian SIRS model in discrete-time lead to an improved goodness of fit. We compare here the results obtained with our model with those provided by a deterministic continuous-time formulation similar to the one implemented in Acedo et al. (2010). As mentioned in the Introduction, deterministic compartmental models in continuous-time are widely used to assess the effect of vaccination programs. By considering only one age-group and a constant population size, the deterministic continuous-time model without vaccines can be formulated as:

$$
\begin{aligned}
\frac{dS}{dt} &= -\beta(t)SI + bR + \mu - \mu S \\
\frac{dI}{dt} &= -aI + \beta(t)SI - \mu I \\
\frac{dR}{dt} &= -bR + aI - \mu R
\end{aligned} \tag{9}
$$

where $\beta(t)$ is defined as $b_0 + b_1 cos(2\pi t + \psi)$ to account for seasonality. The flowchart of this model is shown in Figure 8.

Similar to Acedo et al. (2010), we have assumed that 1% of infants are infected in January 1999 while the remaining 99% of infants are susceptible. We have also set $\mu = 0.01074$, $b = 1.59$, and $a = 36.5$. In order to estimate parameters b_0, b_1 and ψ, we have used the `dsolve` package (Soetaert, Petzoldt and Setzer, 2010) in R (R Core Team, 2017) together with the `optim` function.

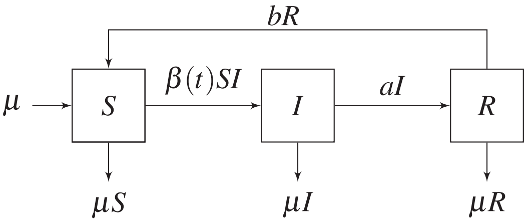


Figure 8: *Flowchart of the deterministic continuous-time SIRS model without vaccination.*

Figure 9 compares the estimates of RSV hospitalizations obtained with both the Bayesian stochastic SIRS model in discrete-time and its deterministic counterpart. As can be seen, the deterministic continuous-time approach is not able to properly describe epidemic peaks. The seasonal pattern is constant over time and it does not explain particular features of annual epidemics. The fitting RMSE are, respectively, 2.52 and 19.09. These results highlight the importance of taking into account the stochasticity inherent in the transmission dynamics.

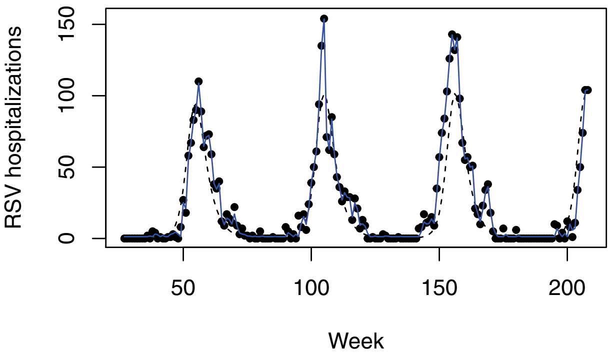


Figure 9: *Real hospitalizations for RSV (solid points) and the estimates obtained with both the Bayesian stochastic model in discrete-time (solid line) and the deterministic continuous-time model (dashed line).*

If we assume that a proportion of newborns are vaccinated, the deterministic model can be reformulated as:

$$
\begin{aligned}
\frac{dS}{dt} &= -\beta(t)SI + bR + (1-\nu)\mu - \mu S \\
\frac{dI}{dt} &= -aI + \beta(t)SI - \mu I \\
\frac{dR}{dt} &= -bR + aI - \mu R \\
\frac{dV}{dt} &= \nu\mu - \mu V
\end{aligned} \tag{10}
$$

Figure 10 displays the real number of hospitalizations from week beginning July 2nd 2001 to week beginning December 20th 2004 and the numbers simulated by the deterministic continuous-time model for a coverage rate $\nu = 0.2$. For comparative purposes, we have also included the results provided by our model. As expected, the deterministic model does not explain properly epidemic peaks. Nevertheless, we can conclude that both strategies show a similar decreasing trend in the number of RSV hospitalizations after the introduction of the vaccine.

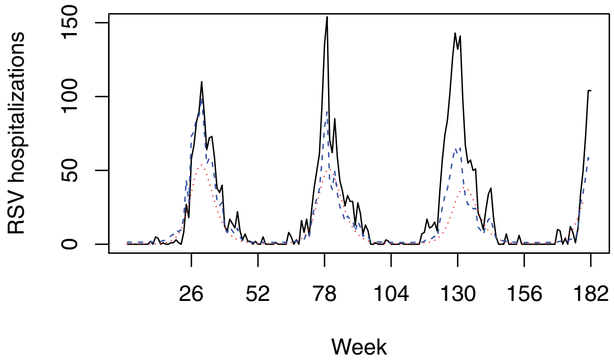


Figure 10: *Real hospitalizations for RSV (solid line) and simulated numbers of hospitalizations provided by both the Bayesian stochastic SIRS model (dashed line) and the deterministic continuous-time model (dotted line) for a coverage rate $\nu = 0.2$.*

5. Conclusion

In this paper, we have described a stochastic compartmental model in discrete-time to describe RSV dynamics in the region of Valencia. However, the model can be adapted for other infectious diseases with (or without) a seasonal pattern and temporary (or permanent) immunity, replacing the transmission rate and the immunity rate according to the nature of disease. Unlike standard formulations, this compartmental model does not assume mass-action mixing of individuals in the population. In addition, the model considers the stochasticity inherent in the transmission of disease and, consequently, it provides a more realistic and accurate description of the progression of infections.

The extended model proposed in this paper provides a useful framework to address one of the important needs in RSV incidence control: the implementation of an efficient vaccination strategy. In particular, we have studied the effects of a vaccination strategy that consists on vaccinating a proportion of newborns, which are the highest priority target population. Additionally, we have pointed out how to adapt the model to simulate a vaccination strategy targeted to pregnant women.

Nevertheless, the model has some limitations. We have assumed that the number of births equals the number of deaths so that the total population size is constant. In addition, we do not consider an age structure into the formulation of the compartmental model. It would be valuable to extend the proposed model to allow for different age groups, for instance infants < 6 months of age and children ≥ 6 months of age, which are considered as distinct target populations for RSV vaccines. An age-structured model would provide an important tool to study the effects of alternative vaccination strategies. It could demonstrate how immunization of a target population may protect others. Besides, this formulation could be used to simulate the benefits of implementing a maternal vaccination strategy followed by direct older infant vaccination as maternal antibody wanes.

Note that we have only implemented a control strategy based on vaccination. However there are other possibilities such as isolation of infected individuals. This alternative control strategy could be straightforwardly incorporated into the model by adapting the probability of becoming infected. Under this scenario, the probability of infection at a particular time point would depend only on a proportion of infected individuals at the previous time point (the ones that have not been isolated). It would also be interesting to assess the impact of both control strategies simultaneously.

Another very fruitful area for further research is the extension of the proposed model to the spatial domain. Space can play a significant role in RSV transmission. In addition, a spatio-temporal model may be useful to detect high-risk areas in need of more strong intervention strategies to reduce the burden of disease.

Finally, it is worth emphasizing that we have focused here on models that have been previously proposed to analyse the impact of vaccination strategies; in particular, we have focused on compartmental models. However, the literature on models for the analysis of infectious disease data is vast and can be found in both statistical as well as epidemiological journals. Comprehensive coverage of statistical models for the analysis of infectious diseases in a single paper is not possible and it is beyond the scope of this paper. Nevetheless, it would be interesting to extend common approaches to model count time series (such as INAR models (Rao and McCabe, 2016) or p-splines (Eilers, Marx and Durban, 2016)) to incorporate the impact of vaccination programs and compare the performance of these different approaches.

Acknowledgement

This work has been supported by Grant Number MTM2014-56233-P from the Spanish Ministry of Economy and Competitiveness.

References

Acedo, L., Díez-Domingo, J., Moraño, J.A. and Villanueva, R.J. (2010). Mathematical modelling of respiratory syncytial virus (RSV): vaccination strategies and budget applications. *Epidemiology and Infection*, 138, 853–860.

Allen, L.J. (2008). An introduction to stochastic epidemic models. In Brauer, F., van den Driessche, P. and Wu, J. (Eds.) *Mathematical Epidemiology*. Springer, Berlin.

Anderson, L.J., Dormitzer, P.R., Nokes, D.J. Rappuoli, R. Roca, A. and Graham, B.S. (2013). Strategic priorities for respiratory syncytial virus (RSV) vaccine development. *Vaccine*, 31S, B209–B215.

Beeler, J.A. and Eichelberger, M.C. (2013). Influenza and respiratory syncytial virus (RSV) vaccines for infants: safety, immunogenicity, and efficacy. *Microbial Pathogenesis*, 55, 9–15.

Bjørnstad, O.T., Finkenstädt, B.F. and Grenfell, B.T. (2002). Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs*, 72, 169–184.

Boys, R.J. and Giles, P.R. (2007). Bayesian inference for stochastic epidemic models with time-inhomogeneous removal rates. *Journal of Mathematical Biology*, 55, 223–247.

Brauer, F. (2008). Compartmental models in epidemiology. In Brauer, F., van den Driessche, P. and Wu, J. (Eds.) *Mathematical Epidemiology*. Springer, Berlin.

Christensen, H., Hickmana, M., Edmunds, W.J. and Trotter, C.L. (2013). Introducing vaccination against serogroup B meningococcal disease: an economic and mathematical modelling study of potential impact. *Vaccine*, 31, 2638–2646.

Contreras, J.R. Vacuna contra el Virus Respiratorio Sincitial (VRS). Asociación Española de Vacunología. Informe Técnico `http://www.vacunas.org/vacuna-contra-el-virus-respiratorio-sincitial-vrs/` [In Spanish. Accessed in: 26th July 2016].

Corberán-Vallet, A. and Santonja, F.J. (2014). A Bayesian SIRS model for the analysis of respiratory syncytial virus in the region of Valencia, Spain. *Biometrical Journal*, 56, 808–818.

Craig, A.P., Hanger, J., Loader, J., Ellis, W.A.H., Callaghan, J., Dexter, A., Jones, D., Beagley, K.W., Timms, P. and Wilson, D.P. (2014). A 5-year Chlamydia vaccination programme could reverse disease-related koala population decline: predictions from a mathematical model using field data. *Vaccine*, 32, 4163–4170.

Dudas, R.A. and Karron, R.A. (1998). Respiratory syncytial virus vaccines. *Clinical Microbiology Reviews*, 11, 430–439.

Eilers, P.H.C., Marx, B.D. and Durban, M. (2015). Twenty years of P-splines. *SORT-Statistics and Operations Research Transactions*, 39, 149–186.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515–533.

Gibson, G.J. and Renshaw, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *IMA Journal of Mathematics Applied in Medicine and Biology*, 15, 19–40.

Higgings, D., Trujillo, C. and Keech, C. (2016). Advances in RSV vaccine research and development - a global agenda. *Vaccine*, 34, 2870–2875.

Hogan, A.B., Glass, K., Moore, H.C. and Anderssen, R.S. (2016). Exploring the dynamics of respiratory syncytial virus (RSV) transmission in children. *Theoretical Population Biology*, 110, 78–85.

Jones, B.G., Sealy, R.E., Surman, S.L., Portner, A., Rusell, C.J., Slobod, K.S., Dormitzer, P.R., DeVincenzo, J. and Hurwitz, J.L. (2014) Sendai virus-based RSV vaccine protects against RSV challenge in an in vivo maternal antibody model. *Vaccine*, 32, 3264–3273.

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.

Ma, Z. and Li, J. (2009). *Dynamical Modelling and Analysis of Epidemics*. World Scientific, London.

O'Neill, P.D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences*, 180, 103–114.

Parra, A., Jiménez, C., Hernández, S., García J.E. and Cardona, A.M. (2013). Bronquiolitis: artículo de revisión. *Neumología Pediátrica*, 8, 95–101.

Pérez-Breva, L., Villanueva R.J., Villanueva-Oller, J., Acedo L., Santonja F.J., et al. (2014). Optimizing strategies for meningococcal C disease vaccination in Valencia (Spain). *BMC Infectious Diseases*, 14, 280.

Poletti, P., Merler, S., Ajelli, M., Manfredi, P., Munywoki, P.K., Nokes, D.J. and Melegaro, A. (2015). Evaluating vaccination strategies for reducing infant respiratory syncytial virus infection in low-income settings. *BMC Medicine*, 13–49.

R Development Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, `http://www.R-project.org`.

Rao, Y. and McCabe, B. (2016). Real-time surveillance for abnormal events: the case of influenza outbreaks. *Statistics in Medicine*, 35, 2206–2220.

Roberts, J.N., Graham, B.S., Karron, R.A., Munoz, F.M., Falsey, A.R., Anderson, L.J., Marshall, V. Kim, S. and Beeler, J.A. (2016). Challenges and opportunities in RSV vaccine development: meeting report from FDA/NIH workshop. *Vaccine*, 34, 4843–4849.

Soetaert, K., Petzoldt, T. and Setzer, R.W. (2010). Solving differential equations in R: Package `deSolve`. *Journal of Statistical Software*, 33, 1–25.

Van Hoek, A.J., Melegarob, A., Zaghenid, E., Edmundsa, W.J. and Gayb, N. (2011). Modelling the impact of a combined varicella and zoster vaccination programme on the epidemiology of varicella zoster virus in England. *Vaccine*, 29, 2411–2420.

Vannia, T., Mendes, P., Fossa, A., Mesa-Friasa, M. and Legooda, R. (2012). Economic modelling assessment of the HPV quadrivalent vaccine in Brazil: a dynamic individual-based approach. *Vaccine*, 30, 4866–4871.

Weidemann, F., Dehnert, M., Koch, J., Wichmann, O. and Höhle, M. (2014). Modelling the epidemiological impact of rotavirus vaccination in Germany – A Bayesian approach. *Vaccine*, 32, 5250–5257.

World Health Organization. *Using the Results of Modelling to Inform WHO Guidelines*. `http://www.who.int/kms/modelling_whoguidelines_2016/en/`

Yu, Z., Liu, J., Wang, X., Zhu, X., Wang, D. and Han, G. (2016). Efficient vaccine distribution based on a hybrid compartmental model. *PLoS One*, 11, e0155416.

SORT 41 (1) January-June 2017, 177-188 DOI: 10.2436/20.8080.02.57

Statistical modelling of warm-spell duration series using hurdle models

Jesper Rydén[*]

Abstract

Regression models for counts could be applied to the earth sciences, for instance when studying trends of extremes of climatological quantities. Hurdle models are modified count models which can be regarded as mixtures of distributions. In this paper, hurdle models are applied to model the sums of lengths of periods of high temperatures. A modification to the common versions presented in the literature is presented, as left truncation as well as a particular treatment of zeros is needed for the problem. The outcome of the model is compared to those of simpler count models.

MSC: 62J12, 62P12.

Keywords: Count data, hurdle models, Poisson regression, negative binomial distribution, climate.

1. Introduction

Regression models for counts arise when the response variable is a count, i.e. a non-negative random number. Often a distribution is specified for the response variable and likelihood-based inference can be performed, with maybe the most common choice for the response being the Poisson distribution, leading to Poisson regression. However, the simpler models are not able to successfully model situations with, for example, excess zeros or truncated observations. Models have been developed in the literature, see Cameron and Trivedi (2013) for a review.

Some statistical problems in the earth sciences are linked to count data. In particular, regression models for count data could be of interest, when facing series of so-called climate indicators (occasionally called climate indices). These are often numbers relating to extreme phenomena, for instance *heat waves* or *warm spells*, loosely described

[*] Department of Mathematics. Uppsala University. Box 480. SE 751 06 Uppsala. Sweden.
Office phone: +46 18 4713288. E mail: jesper.ryden@math.uu.se

Received: March 2016
Accepted: February 2017

as periods of unusually hot weather. In the literature, the notion of a heat wave is often reserved for periods of great severity, for instance causing deaths among people. For typical Swedish conditions, analysed in the sequel, the notion warm spell is therefore preferred. In climatology, interest concerns changes in frequency, intensity or duration of such quantities.

From a data-analytic point of view, a warm spell is a *run*, i.e. a period of consecutive days when the maximum is above a specified high value. In this paper, we examine statistical modelling of the indicator *warm-spell duration index* (WSDI), defined as the annual count of days with at least 6 consecutive days when the daily maximum temperature is exceeding a predefined threshold (see exact definition in the sequel). However, the statistical modelling of such sequences imply several challenges. In this paper, we focus on the fact that observations are truncated, but in addition, an annual count of zero might also be observed depending on the location and its occasionally cold climate. In fact, such models for count data seem not to have been studied in the applied literature, either in climate research or other applications.

A key issue in climatology research is investigation of trends. A methodology for count data could be to check independence, and if possible, use time as a covariate in a regression model. Generalised linear models and their extensions are then natural candidates for modelling. As stated by Chandler and Scott (2011), applications of such models in environmental trend analysis have so far been relatively limited. Examples are rare, but similar statistical concepts are found for instance in Frei and Schär (2001), a recent study on extreme precipitation was made by Hertig et al (2014), and trends of flash counts are discussed by Bates, Chandler and Dowdy (2015). Concerning an indicator related to the annual number of warm spells, Rydén (2015) investigated a possible trend for the city of Uppsala, Sweden. Then the elements of the time series were simply non-negative integers, and the Poisson distribution was found to be a reasonable description. Moreover, the sequence was considered independent, and hence Poisson regression was applied.

The paper is organised as follows. In the next section, the indicator WSDI is defined and discussed, along with a presentation of the source of the data. In Section 3, the framework of hurdle models is introduced, including the modification needed for modelling of the WSDI. In Section 4, data are introduced and the results of applying the hurdle models are presented, and finally in Section 5, a summary and discussion is given.

2. Warm-spell duration index

Several indicators, also labelled indices, have been suggested for monitoring change in climatic extremes (see e.g. Frich et al 2002), but as pointed out by Perkins and Alexander (2013) concerning heat waves: "Clear and common definitions, at least for some types of extreme events, remain rare and nonexistent". Climate indices may relate to temperatures as well as precipitation, they may be based on absolute thresholds or percentile based. Thus, definitions have to be clearly stated in research work. An overview of indices, as well as results from an analysis of trends at a global level, is given by Alexander et al (2006).

Data were retrieved online from the website of the European Climate Assessment & Dataset (ECA&D) project[1]. Definitions of indices are found at the webpage of the joint CCl/CLIVAR/JCOMM Expert Team (ET) on Climate Change Detection and Indices (ETCCDI)[2].

The indicator WSDI, warm-spell duration index, belongs to the category of duration indices. Such indices define periods of excessive warmth, cold, wetness or dryness. WSDI is defined as the annual count of days with at least 6 consecutive days when the daily maximum temperature is exceeding the threshold T_{90}. To be more precise: Let $T(i,j)$ be the daily maximum temperature on day i in year j and let T_{90} be the calendar day 90th percentile, centred on a five-day window for the base period 1961-1990. Then the number of days per year j is summed where, in intervals of at least 6 consecutive days, $T(i,j) > T_{90}$.

Note that the annual count, the annual observation of WSDI is a sum of all days belonging to a warm-spell period. The number of warm spells is not taken into account, so a year with two spells of lengths 6 and 8 days, respectively, would result in a value of WSDI equal to 14, the same value as a year with a single long spell of 14 days.

3. Hurdle models for count data

In this section we review hurdle models for count data (cf. Winkelmann 2008, Cameron and Trivedi 2013), and discuss implications for the application introduced previously and possible alternatives for the modelling.

3.1. Structure of the hurdle-count model

We commence by recalling the notion of a truncated random variable. Consider a random variable Y, defined on $0,1,2,\ldots$. Now assume that only values $y > a$ are observed. The truncated distribution $\widetilde{Y}$ then has the probability-mass function

$$p_{\widetilde{Y}}(\widetilde{y}) = \frac{1}{1-F_Y(a)} p_Y(\widetilde{y}), \quad \widetilde{y} = a+1, a+2, \ldots \ .$$

[1] http://www.ecad.eu/

[2] http://etccdi.pacificclimate.org/list_27_indices.shtml

With two-part models for counts, a model is introduced where the probabilistic properties of zero counts differ from other (positive) counts. Such models were proposed by Mullahy (1986). For a random variable Y, suppose that we observe either $Y = 0$ or $Y > a$. For the zero component, we introduce the probability-mass function $p_1(y)$ and for the positive outcomes, we consider the (unrestricted) probability-mass function $p_2(y)$; related distribution functions are $F_1(y)$ and $F_2(y)$. A *hurdle model* is then defined by

$$\mathsf{P}(Y = j) = \begin{cases} p_1(0) & \text{if } j = 0 \\ \frac{1-p_1(0)}{1-F_2(a)} p_2(j) & \text{if } j > a \end{cases} \tag{1}$$

(For $0 < j \leq a$, the probability-mass function takes the value zero.) Defining a binary, censoring indicator

$$d = \begin{cases} 1, & \text{if } y > a \\ 0, & \text{if } y = 0 \end{cases}$$

the probability-mass function for an outcome y with indicator d can then be written as

$$\begin{aligned} p(y) &= p_1(0)^{1-d} \left[\frac{1-p_1(0)}{1-F_2(a)} p_2(y) \right]^d \\ &= \left[p_1(0)^{1-d} (1-p_1(0))^d \right] \left[\frac{p_2(y)}{1-F_2(a)} \right]^d . \end{aligned}$$

3.2. Estimation

We now turn to estimation. In a regression context, suppose we have a covariate x. Introducing parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, the probability functions can be notated $p_1(y; x, \boldsymbol{\theta}_1)$ and $p_2(y; x, \boldsymbol{\theta}_2)$.

The log-likelihood function then follows, with observations $(x_1, y_1), \dots, (x_n, y_n)$, as

$$\begin{aligned} \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \ell_1(\boldsymbol{\theta}_1) + \ell_2(\boldsymbol{\theta}_2) \\ &= \sum_{i=1}^{n} \left[(1-d_i) \ln p_1(0; x_i, \boldsymbol{\theta}_1) + d_i \ln(1 - p_1(0; x_i, \boldsymbol{\theta}_1)) \right] \\ &\quad + \sum_{i=1}^{n} d_i \left[\ln p_2(y_i; x_i, \boldsymbol{\theta}_2) - \ln(1 - F_2(a; x_i, \boldsymbol{\theta}_2)) \right] . \end{aligned} \tag{2}$$

Thus, the log-likelihood function can be maximised by separately maximising each component, which certainly simplifies the numerical treatment.

3.3. Specification of count distributions

Many options obviously exist for choosing the distributions $p_1(.)$ and $p_2(.)$. In the original paper by Mullahy (1986), these were specified to be of the same family. Common practice now is to specify different processes for $p_1(.)$ and $p_2(.)$. The binary process, $p_1(.)$, is often modelled as a logit model, while $p_2(.)$ is chosen as a Poisson or negative binomial distribution. After preliminary analysis of data, overdispersion was found present and truly significant (p value $1.6 \cdot 10^{-4}$, test by Cameron and Trivedi (1990), as implemented in the routine `dispersiontest` in the R package `AER`, see Kleiber and Zeileis, 2008). Thus a negative binomial distribution was applied, and will be discussed next.

Several characterisations of the negative binomial distribution exist, in terms of parameterisation, and we chose in this work to employ the distribution with probability-mass function as follows, the so-called Negbin II:

$$p(z;\mu,\alpha) = \frac{\Gamma(\alpha^{-1}+z)}{\Gamma(\alpha^{-1})\Gamma(z+1)} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{1/\alpha} \left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{z}, \quad z = 0,1,2,\ldots \tag{3}$$

where in a regression context with a covariate x,

$$\mu = \exp(\beta_0 + \beta_1 x)$$

and α is a dispersion parameter. When $\alpha \to 0$, the Poisson distribution is obtained as a limit. Moreover,

$$\mathsf{E}[Z] = \mu, \qquad \mathsf{V}[Z] = \mu(1+\alpha\mu).$$

In summary; with regards to the likelihood estimation in our problem, we have the parameter vector $\boldsymbol{\theta}_2 = (\beta_0,\ \beta_1,\ \alpha)$. For the binary part,

$$p_1(y,x_i,\boldsymbol{\theta}_1) = \frac{\exp(\beta_0' + \beta_1' x_i)}{1+\exp(\beta_0' + \beta_1' x_i)}$$

and thus $\boldsymbol{\theta}_1 = (\beta_0',\ \beta_1')$.

Remark. In most texts, and computer implementations (in R, e.g. Zeileis, Kleiber and Jackman, 2008), hurdle models with $a = 0$ in Eq. (1) are considered; that is, the hurdle separates zeros from positive observations. In our application, we have the possible outcomes $0,6,7,8,\ldots$, and hence $a = 5$ in Eq. (1); to the author's knowledge, this is a situation rarely met in applications considered in the literature. In Stata, modelling with truncated hurdle models is implemented. An example with a truncated Poisson distribution for the non-zero counts is given by McDowell (2003).

3.4. Mean for the hurdle-count model

For the hurdle model in Eq. (1), for the sake of notation, introduce

$$b = \frac{1 - p_1(0)}{1 - F_2(a)}. \tag{4}$$

Moments about the origin then follow as

$$\mathsf{E}[Y^k] = 0^k p_1(0) + \sum_{y=a+1}^{\infty} y^k b\, p_2(y) = b \sum_{y=a+1}^{\infty} y^k p_2(y). \tag{5}$$

Cameron and Trivedi (2013), Section 4.12, give the corresponding derivation for the case where $a = 0$, and the resulting formula can then be expressed in terms of the expected value for the distribution $F_2(y)$.

For model diagnostics (in our application, see Section 4.3), we may use $\mathsf{E}[Y]$ following Eq. (5), plugging in estimates. Then assuming a logit model for the binary part, we find estimates for the quantities in the factor b in Eq. (4):

$$\widehat{p_1(0)} = p_1(0; x_i, \widehat{\boldsymbol{\theta}}_1) = \frac{\exp(\widehat{\beta}'_0 + \widehat{\beta}'_1 x_i)}{1 + \exp(\widehat{\beta}'_0 + \widehat{\beta}'_1 x_i)}$$

and $\widehat{F_2(a)} = F_2(a, x_i, \widehat{\boldsymbol{\theta}}_2)$, where $F_2(.)$ is the related distribution for the Negbin II distribution.

4. Modelling warm-spell duration index

We have chosen to investigate time series of annual observations of WSDI from three locations and periods in Sweden: Falun (60°37′N, 15°37′E), 1914-2010, Stockholm (59°21′N, 18°03′E), 1914-2014, and Uppsala (59°51′N, 17°37′E), 1914-2011. This initial choice was made based on data quality (quite long series without gaps, though missing data for 2006 at Falun) and we have, moreover, two locations quite close in distance (Stockholm and Uppsala, less than 10 km).

4.1. Dependence issues

In order to apply regression models for counts using time as a covariate, dependence in each of the sequences was first investigated, by checking plots of autocorrelation functions and performing the Ljung–Box test of independence in time series (Ljung and Box, 1978). Consider a time series $x_1, \ldots, x_N$. The null hypothesis is here that the first

m autocorrelations are jointly zero:

$$H_0 : \quad \rho_1 = \rho_2 = \cdots = \rho_m = 0.$$

Let r_k be the sample autocorrelation function at lag k and m the number of lags being tested. The test statistic is given by

$$Q = N(N+2) \sum_{k=1}^{m} \frac{r_k^2}{N-k}$$

which under the null hypothesis is distributed as $Q \sim \chi^2(m)$. A choice of m has to be made; in the literature, it has been suggested that $m \approx \ln N$ (Tsay, 2010). For our locations we find Falun ($p = 0.85$), Stockholm ($p = 4.3 \cdot 10^{-10}$), Uppsala ($p = 6.2 \cdot 10^{-8}$). For the two last locations, we thus reject the null hypothesis about independence. Plots of empirical autocorrelation functions strengthen this result. Thus, a more evolved time-series model for counts would have to be introduced for these two locations. One option for further modelling could be to introduce a time-series model for counts, for instance, of the class INAR (Al-Osh and Alzaid, 1987; Alzaid and Al-Osh, 1990; for a recent review, see Scotto, Weiss and Gouveia, 2015).

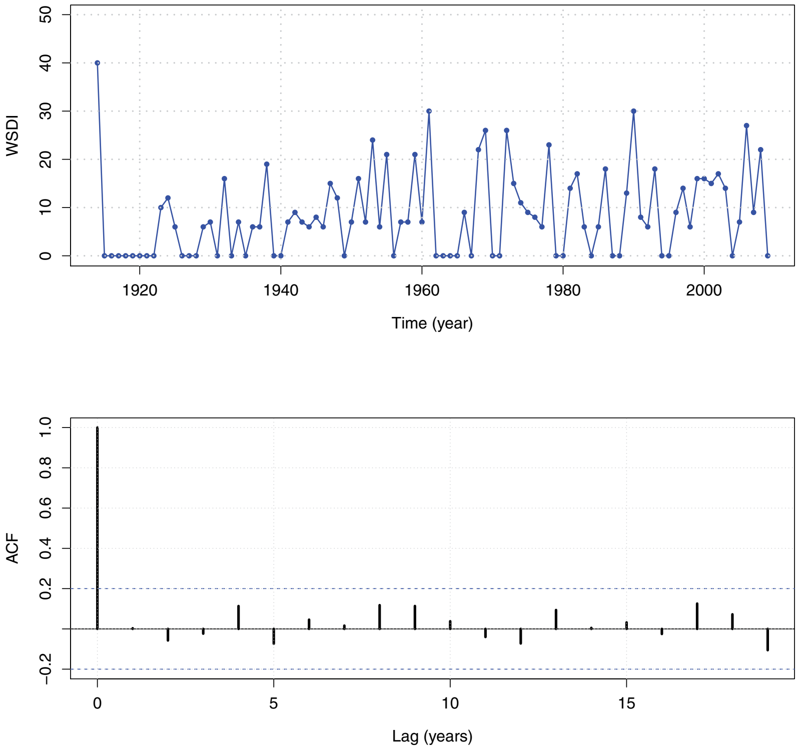


Figure 1: *Top: Time series of WSDI at Falun. Bottom: Sample autocorrelation function of WSDI at Falun.*

For the location of Falun, we choose to continue, using time as a covariate in a regression model. The original time series[3] at that location is shown in Figure 1, top panel. The proportion of zeros is found to be 0.35. The sample autocorrelation function is displayed in Figure 1, bottom panel, and we conclude that data could be considered a sequence of independent observations.

4.2. Likelihood inference

For the zero part, i.e. maximisation of the function $\ell_1(\boldsymbol{\theta}_1)$ in Eq. (2), a logistic regression was performed with a binary response variable $\pi(x)$ and the related model

$$g(x) = \beta_0' + \beta_1' x$$

where $g(x) = \ln(\pi(x)/(1-\pi(x)))$ and the covariate x indicates time. Hence, the vector $\boldsymbol{\theta}_1 = (\beta_0' \ \beta_1')$. The routine `glm` in the statistical software package R (R Core Team 2016) was employed. The estimation procedure resulted in estimates $\hat{\beta}_0' = -0.25$ and $\hat{\beta}_1' = 0.018$ with related p-values 0.56 and 0.025, respectively. The covariate time is thus significant.

For the maximisation of the log-likelihood function $\ell_2(\boldsymbol{\theta}_2)$, a Negbin II was assumed (see Eq. (3)). The optimisation was carried out by the routine `optim`, using the procedure by Nelder and Mead (1965). The following point estimates, with related standard errors within parentheses as obtained from the inverted observed Fisher information matrix, were obtained:

$$\hat{\beta}_0 = 1.01 \ (1.70), \quad \hat{\beta}_1 = 0.0053 \ (0.0062), \quad \alpha = 2.84 \ (5.98)$$

with p-values 0.56, 0.39 and 0.64 respectively. With the climate application in focus, we note that the slope is slightly positive in magnitude, and not statistically significant.

4.3. Model checking and comparison

In Figure 2, the original time series is plotted along with the mean of the fitted model, following Eq. (5).

We deduced earlier that the original time series could be considered an independent sequence (cf. Figure 1, bottom panel). In Figure 3, the sample autocorrelation function of the raw residuals is shown. We note that dependence is still not a concern.

[3] The high observed value 40 at the beginning of the series belongs to year 1914, the summer of which is known for historians as being unusually warm and pleasant, right before the outburst of World War I, end of July 1914.

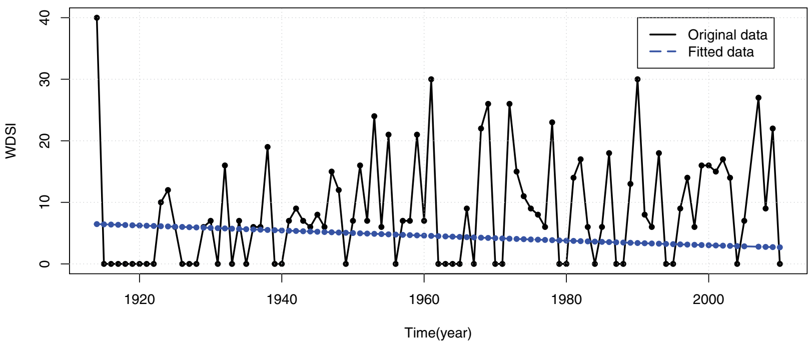


Figure 2: *Original time series and fitted model.*

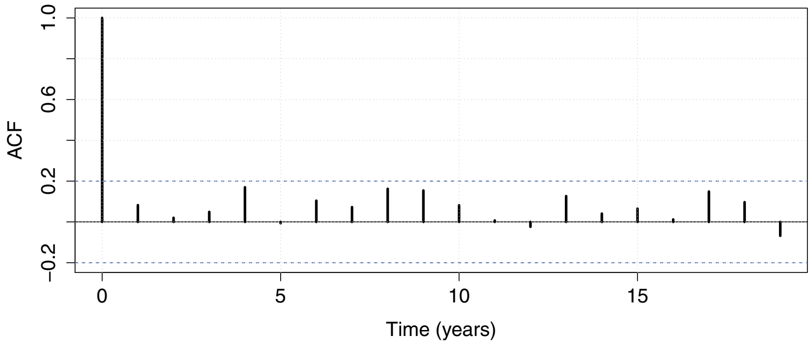


Figure 3: *Sample autocorrelation function for the residuals.*

One might contemplate a simpler statistical model, though not taking into account the particular structure of data. A negative binomial distribution with mean $\mu = \exp(\beta_0 + \beta_1 x)$, say, could then be directly fitted to all outcomes of WSDI in a regression model (using the routine `glm.nb` in the package MASS in R). Such a model results in an estimate $\hat{\beta}_1 = 0.0092$ for the slope (standard error 0.0055) with the related p-value 0.092, which could be compared to the corresponding estimate for the hurdle model.

For model comparison, values of AIC (Akaike's Information Criterion, Akaike 1973) are useful:

$$\text{AIC} = 2k - 2\ln L$$

where k is the number of parameters and L the value at the optimum of the likelihood function. These were computed for the two considered models. For the hurdle model, AIC = 505.7, while for the approach with negative binomial, AIC = 597.8. A model with as small AIC as possible is preferable, and there is hence some merit of the hurdle model in this respect.

5. Discussion

Regression models for counts find applications in many scientific fields. Typically, a Poisson distribution is assumed for count data, but the original models have to be modified in order to model e.g. overdispersion or excesses of zeros. One special model is the so-called hurdle model, attributed to Mullahy (1986). As a further, quite recent, development of the hurdle model could be mentioned Saffari, Adnan and Greene (2012), where a framework with hurdle models adopted to right-censored data was presented (application to counts of fish).

In this paper, a hurdle model for the case of left-truncated data was presented, motivated by an application from climatology where data is either zero or an integer at least six. Estimation was carried out by likelihood techniques. The obtained results were compared with estimates from a simpler model, fitting a negative binomial distribution directly to the counts. Point estimates of trend (coefficient for slope) became of roughly the same magnitude. Comparison of AIC indicates that the hurdle model is preferable. However, for all estimated parameters, uncertainties are considerably high, as can be reflected from related p-values.

The meaning of the quantity WSDI as an additive measure of days may have influences on the distribution over possible integers. For instance, a WSDI of 11 can be obtained only as a single period of 11 days, while an observed count of 14 can result in three ways: a single period of 14, adding 6 and 8 or adding 7 and 7. Thus, in addition to natural variability, results could vary due to combinatoric reasons. For the data sets, the following table of counts for various WSDI can be compiled (for Falun, also cf. Figure 1, top panel):

WSDI	7	8	9	10	11	12	13	14	15
Counts, Falun	10	3	5	1	1	2	1	3	3
Counts, Uppsala	6	4	3	1	2	4	6	6	2
Count, Stockholm	6	5	4	1	0	2	3	2	4

We note that for WSDI equal to 10 and 11, few counts are found (not likely combinations to occur). Thus, to model WSDI with a probability distribution, possibly this phenomenon could be taken into account.

For all types of regression models, model assessment is an important objective. A review for the common cases of regression with count data is given by Cameron and Trivedi (2013), where it is also stated in Chapter 5 (Model Validation and Testing) that there is "...considerable scope for generalization and application to a broader range of count data models." In this paper, we made a simple investigation (see Figures 2 and 3). Further research would be to, for instance, develop and examine goodness-of-fit tests for the hurdle model with truncated observations for the non-zero part.

In this paper, regression models for counts were modelled using time as a covariate. It could be mentioned that a non-parametric regression methodology based on P-splines might be a useful approach, see the recent paper by Eilers, Marx and Durbán (2015) in this journal.

Acknowledgements

I am grateful to Dr Hans Bergström, Dept. of Earth Sciences at Uppsala University, for discussion about climate indicators. Thanks also to the referees, whose constructive comments led to improvement of the paper.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Petro, B.V. and Csaki,F., eds., *Second International Symposium on Information Theory*, 267-281, Budapest, Akademiai Kaido.

Alexander, L.V., Zhang, X., Peterson, T.C., Caesar, J., Gleason, B., Klein Tank, A.M.G., Haylock, M., Collins, D., Trewin, B., Rahimzadeh, F., Tagipour, A., Ambenje, P., Rupa Kumar, K., Revadekar, J. and Griffiths. G. (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research*, 111, 1–22.

Al-Osh, M.A. and Alzaid, A.A. (1987). First order integer valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis*, 8, 261–275.

Alzaid, A.A. and Al-Osh, M. (1990). An integer-valued pth-order autoregressive structure (INAR(p)) process. *Journal of Applied Probability*, 27, 314–324.

Bates, B.C., Chandler, R.E. and Dowdy, A.J. (2015). Estimating trends and seasonality in Australian monthly lightning flash counts. *Journal of Geophysical Research: Atmospheres*, 120, 3973–3983.

Cameron, A.C. and Trivedi, P.K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics*, 46, 347-364.

Cameron, A.C. and Trivedi, P.K. (2013). *Regression Analysis of Count Data*. 2nd ed. Cambridge University Press.

Chandler, R.E. and Scott, E.M. (2011). *Statistical Methods for Trend Detection and Analysis in the Environmental Sciences*. Wiley.

Eilers, P.H.C., Marx, B.D. and Durbán, M. (2015). Twenty years of P-splines. *SORT-Statistics and Operations Research Transactions*, 39, 149–186.

Frei, C. and Schär, C. (2001). Detection probability of trends in rare events: theory and applications to heavy precipitation in the Alpine region. *Journal of Climate*, 14, 1568–1584.

Frich, P., Alexander, L.V., Della-Marta, P., Gleason, B., Haylock, M., Klein Tank, A.M.G. and Peterson, T. (2002). Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Research*, 19, 193–212.

Hertig, E., Seubert, S., Paxian, A., Vogt, G., Paeth, H. and Jacobeit, J. (2014). Statistical modelling of extreme precipitation indices for the Mediterranean area under future climate change. *International Journal of Climatology*, 34, 1132–1156.

Kleiber, C. and Zeileis, A. (2008). *Applied Econometrics with R*. Springer-Verlag.

Ljung, G.M. and Box, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297–303.

McDowell, A. (2003). From the help desk: hurdle models. *The Stata Journal*, 3, 178–184.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 341–365.

Nelder, J.A. and Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal*, 7, 308–313.

Perkins, S.E. and Alexander L.V. (2013). On the measurement of heat waves. *Journal of Climate*, 26, 4500–4517.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www-R-project.org/

Rydén, J. (2015). A statistical analysis of trends for warm and cold spells in Uppsala by means of counts. *Geografiska Annaler: Series A, Physical Geography*, 97, 431–436.

Saffari, S.E., Adnan, R. and Greene, W. (2012). Hurdle negative binomial regression model with right censored count data. *SORT – Statistics and Operations Research Transactions*, 36, 181–194.

Scotto, M.G., Weiss, C.H. and Gouveia, S. (2015). Thinning-based models in the analysis of integer-values time series: a review. *Statistical Modeling*, 15, 590–618.

Tsay, R.S. (2010). *Analysis of Financial Time Series*. 3rd ed. Hoboken, NJ: John Wiley & Sons.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. 5th ed. Springer-Verlag

Zeileis, A., Kleiber, C. and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, 27, 1–25.